Der Cloze Test als Meßinstrument der Fremdsprachenkompetenz von Schülern mit Deutsch als Muttersprache

(eine empirische Studie)

Klaus Balik

1981

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41535726_0007

DER CLOZE TEST ALS MESSINSTRUMENT DER FREMDSPRACHEN-KOMPETENZ VON SCHÜLERN MIT DEUTSCH ALS MUTTERSPRACHE (EINE EMPIRISCHE STUDIE)

Inaugural-Dissertation
zur
Erlangung der Doktorwürde
der
Philosophischen Fakultät
der
Ludwig-Maximilians-Universität München

vorgelegt
von
Klaus Balik
aus
München

München 1980

Referent: Prof. Dr. Karlheinz Hecht
Korreferent: Prof. Dr. Leonhard Lipka
Tag der mündlichen Prüfung: 23. 2. 1981

UNIV.
LEIDEN
BIBL.

INHALTSÜBERSICHT

VORWORT 5

1. EINFÜHRUNG 6

1.1. Vorbemerkungen 6
1.1.1. Ziele und Begründung des Forschungsvorhabens 6
1.1.2. Forschungsfragen 9

1.2. Definition und Umfang des Forschungsgegenstandes 9
1.2.1. Definition des Cloze Tests 9
1.2.2. Umfang des Forschungsgegenstandes und Eingrenzung 10
1.2.3. Standardmodell des Cloze Tests 11

1.3. Bearbeitungsschritte - Forschungsmethode 12

1.4. Umfang der Studie 13
1.4.1. Material 13
1.4.2. Eingrenzung der Studie 13
1.4.3. Weiterentwicklung der Studie 13
1.4.4. Besondere Anmerkungen 14

2. VORGEFUNDENE VORAUSSETZUNGEN IN DER LITERATUR UND FORSCHUNG 15

2.1. Grundlagen des Fremdsprachentestens: Intentionalität/Qualität/Modalität 15
2.1.1. Intentionalität 15
2.1.2. Qualitätskriterien von Tests 24
2.1.2.1. Allgemeine Übersicht 24
2.1.2.2. Objektivität 26
2.1.2.3. Reliabilität 28
2.1.2.4. Validität 35

2.1.2.5. Normierung 40
2.1.2.6. Ökonomie und Testadministration 41
2.1.2.7. Wechselbeziehungen zwischen den Gütekriterien 42
2.1.3. Modalität des Fremdsprachentestens 45
2.1.3.1. Geschichtliche Einführung: drei Phasen der Entwicklung 45
2.1.3.2. Der vorwissenschaftliche Trend 45
2.1.3.3. Der strukturalistisch-psychometrische Trend 46
2.1.3.4. Der integrativ-soziolinguistische Trend 51

2.2. Die Fremdsprachenkompetenz 53

2.3. Vorgefundene schulpädagogische und didaktische Grundlagen in der Sekundarstufe I 59

2.4. Der Cloze Test 65
2.4.1. Entwicklung und Hintergrund 65
2.4.2. Forschung zum Stand des Cloze Tests als Meßinstrument der Fremdsprachenkompetenz 73
2.4.3. Forschung zu methodischen Aspekten des Cloze Tests 95
2.4.4. Theoretische Grundlagen des Cloze Tests 107
2.4.4.1. Empirische Antworten auf theoretische Grundfragen 107
2.4.4.2. Betrachtungen der Funktionsweisen des Cloze Tests 109

3. DISKUSSION UND PROBLEMATIK 115

4. METHODE 121

4.1. Vorversuche 121

4.2. Forschungshypothesen 127

4.3. Plan zur Durchführung der Studie

4.4. Die Probanden 128

4.5. Die Testkonstruktion 131

4.6. Die Erhebung der Validierungskriterien 133

4.7. Die Testdurchführung 133

5. RESULTATE 136

5.1. Mittelwerte, Schwierigkeitsindizes und Streuungswerte 136

5.1.1. Arithmetische Mittelwerte 137
5.1.2. Indizes der Textschwierigkeit 140
5.1.3. Verteilung der Rohpunkte 145
5.1.4. Die Mittelwerte der Validierungskriterien 152

5.2. Reliabilität 155

5.3. Validität 160

5.4. Ökonomie und Testadministration 164

5.5. Interdependenz der statistischen Werte 166

5.6. Ergänzende Beobachtungen 172

6. ZUSAMMENFASSUNG UND FOLGERUNGEN 175

7. APPENDIX 182

7.1. Testbeispiele mit Lösungsblättern 182

7.2. Gesamtübersicht der statistischen Ergebnisse der Durchführung einer Testbatterie bei Schülern der Jahrgangsstufen 5 bis 9 der Haupt-, Realschulen und Gymnasien 190

8. LITERATURVERZEICHNIS 193

9. VERZEICHNIS DER TAFLEN UND ABBILDUNGEN 212

VORWORT

In der Absicht, einen produktiven und wertvollen Forschungsbeitrag für die Fremdsprachenmethodik zu schaffen, der sowohl dem Lehrer als auch dem Schüler direkt zugute kommt, wurde ich im Anschluß an erste Begegnungen in den Jahren 1970 bis 1974 mit der Testmethodik in den USA durch die Anregung von Prof. Dr. Hecht, dem Lehrstuhlinhaber für die Didaktik der Englischen Sprache und Literatur an der Ludwigs-Maximilians-Universität München, wieder mit dieser Thematik zusammengebracht. Herrn Prof. Hecht ist deshalb hier an erster Stelle mein herzlichster Dank gesagt auch für seine Förderung der Arbeit.

Bei den ersten Vorstößen in das hiesige Neuland der integrativen Fremdsprachentestmethodik war noch nicht absehbar, wie weit mich die Arbeit führen würde. Ohne die Mithilfe vieler Spezialisten, denen ich großen Dank schulde, wäre die Untersuchung kaum durchführbar gewesen. So führte mich die Arbeit wieder zurück nach Los Angeles, wo mir große Hilfe zuteil wurde bei der Literaturbeschaffung und fachlicher Rat von meinen Lehrern und Freunden Dr. Brière, Dr. Brown und Dr. Hinofotis.

Wertvolle Unterstützung bei der statistischen Bearbeitung erhielt ich von Dr. Wiseman vom Leibnitz Rechenzentrum der Universität München, von Dr. Steinack und Dr. Gigerenzer von der Universität München.

Über die Situation der Cloze Test - Forschung in Deutschland informierten mich Dr. James von der Universität Erlangen und Dr. Stevenson von der Gesamthochschule in Essen.

Vor allem wäre ohne die Mitwirkung der Lehrer Block, Dumsky, Friese, Hirschnagel, Meyer, Pokorny, Quasnitzka, Schmiedel, Schuh, Strobel, Waas, Winckhler sowie Wolff und ohne die Genehmigung des Bayerischen Kultusministeriums die Durchführung der Untersuchung nicht möglich gewesen. Ihnen allen gilt mein Dank.

1. EINFÜHRUNG

1.1. Vorbemerkungen

1.1.1. Ziele und Begründung des Forschungsvorhabens

Der Cloze Test (im folgenden abgekürzt als CT) steht zur Zeit im Mittelpunkt der internationalen Diskussion zum Thema des Fremdsprachentests (siehe GAL-Tagung[1] in Mainz 1978; TESOL Convention[2] in Boston, Februar 1979, und Konferenz der Interuniversitären Sprachtestgruppe 1978 in Essen, 1979 in Bonn). Das vorhandene Forschungsmaterial ist bereits in mehreren Bibliographien (Jongsma, 1971; Robinson, 1972; Riley, 1973; Oller, 1975; Boyce, 1976) gesammelt und nimmt ständig zu. Eine weitere Bibliographie als 'print-out' des ERIC Systems[3] ist geplant.

Oller (1975) ermuntert zum Weiterforschen. Er weist darauf hin, daß trotz des Umfanges seiner Bibliographie "the technique has been vastly underexploited to date". Dies trifft besonders zu für den Bereich unseres mehrgliedrigen Schulsystems, auch für Schüler niedriger Fremdsprachenkompetenzstufen. Erste exploratorische Versuche seit ca. 1977 am Lehrstuhl für Didaktik der englischen Sprache der Universität München, von wo aus auch der Anstoß für die vorliegende Arbeit kam, bewegten sich in völligem Neuland.

Die Attraktivität des Cloze-Verfahrens beruht auf mehreren Faktoren. Es wird einmal argumentiert, daß mit seiner Hilfe das Interimsprachsystem (Interlanguage) des Fremdsprachenlernenden in einer Weise diagnostiziert werden kann, die weitgehend seinem Lernprozeß entspricht bzw. daraus erwächst. Zum anderen wird der CT - allerdings nicht unumstritten, als objektives, schnelles und praktisches Meßinstrument einer allgemeinen Fremdsprachenkompetenz gesehen (Oller, 1973
Im ersteren Sinne wird der Test bereits seit langem bei uns verwen-

---

1 Gesellschaft für Angewandte Linguistik

2 Teaching of English to Speakers of other Languages

3 ERIC Clearinghouse on Languages and Linguistics, Center for Applied Linguistics, Arlington Va.

det und zwar als Einsetztest, der spezifische Ziele des Unterrichtens durch entsprechende Lücken prüft (z.B. Hecht, 1977:57; Kamratowski, 1973: 13). So können in einem Text z.B. das neugelernte Vokabular, bestimmte Verbformen, Präpositionen oder ähnliches ausgelassen werden. Diese Art des Tests, von Valette (1967:212 ff) als 'achievement test' klassifiziert, die als Sonderformen des CT angesehen werden kann, wird hier ausgeklammert. So trennen auch Taylor (bereits 1953) und Coe (1980) ausdrücklich diese Sonderformen vom sogenannten Standard oder Classic Cloze Test, wie er noch ausführlich dargestellt wird (2.4.1.), und bezeichnen sie als 'sentence-completion test' bzw. 'gapfilling test'.
Der zweite Aspekt, das ist der CT als Meßinstrument einer allgemeinen Fremdsprachenkompetenz in seiner Anwendbarkeit in unserem Schulsystem, ist Gegenstand der vorliegenden Arbeit.

Die theoretische Frage nach der Fremdsprachenkompetenz und wie sie erfaßt werden kann, bildet einen Schwerpunkt im zweiten Kapitel. Trotzdem liegt das Hauptgewicht der Studie auf der empirischen Validierung des CT für unser Schulsystem. Da der Englischunterricht der Sekundarstufe I der gesamten BRD im wesentlichen die gleichen schulischen Grundlagen und unterrichtlichen Leitziele hat, wird im folgenden ganz allgemein darauf Bezug genommen, wenngleich die Stichprobenentnahme aus praktischen Gründen nur in Bayern erfolgte. Regionale Unterschiede bleiben unberücksichtigt.

Trotz des Umfanges der bereits geleisteten Forschungsarbeit sowohl zu allgemeinen Grundfragen (z.B. Allen, 1968; Allen & Davies, 1977; Allen & Campbell, 1972; Freedle & Carroll, 1972; Oller, 1973 und 1979; Spolsky, 1973; Taylor, 1953), als auch spezieller Studien zum 'standard cloze' (cf. 1.2.3.) im fremdsprachlichen Bereich (z.B. Anderson, 1978 und 1970; Darnell, 1970; Hinofotis, 1976; Irvine/Atai/Oller, 1974; Krzyzanowsky, 1976; Oller, 1972; Oller/Conrad, 1971; Pike, 1973 und Wijnstra & Wageningen, 1974) liegen noch keine Untersuchungen zur Verwendbarkeit des CT in unserem Schulsystem und speziell für niedrige Kompetenzstufen vor. Die vorhandenen Ergebnisse können auch nicht einfach auf unser System übertragen werden, denn die verwendeten Daten stammen von Sprechern anderer Muttersprachen, mit anderen Lernhintergründen (z.B. verschiedene Lernziele, soziokulturelle Bedingungen, Methoden etc.) und zum Teil auch anderen Zielsprachen.
Nachweise hiefür liefern z.B. die Untersuchungen von Pike (1973) und Hisama (1976). Deshalb kann diese Studie beim Vergleich der internationalen Ergebnisse einen Beitrag zu einer umfassenderen Validierung des CT leisten.

Weitere Gründe, die eine Untersuchung der Anwendbarkeit des CT in unserer spezifischen Fremdsprachensituation rechtfertigen, sind:

a) Die Neuorientierung des Englischunterrichts auf das Leitziel der Kommunikativen Kompetenz (siehe 2.2.); dies erfordert entsprechende Meßinstrumente. Man kann argumentieren, daß der CT wegen seiner Ähnlichkeit mit Kommunikationssituationen (Sinnentnahme aus Texten, siehe 2.4.4.2.) ein gültiges Meßinstrument der kommunikativen Kompetenz ist. Es ist fraglich, ob 'non-native' Lehrer ohne entsprechende Hilfen überhaupt in der Lage sind, kommunikative Wirksamkeit und die Angemessenheit von Schüleräußerungen [4a] zu prüfen (siehe Projekt Hecht, 1980/81).

b) Der Fremdsprachenunterricht nimmt eine wichtige Stellung ein, wenn es darum geht, für den Schüler die ihm gemäße Schulart zu finden (z.B. Schüler, die den Fremdsprachenstandard eines Gymnasiums nicht erreichen, werden an die Real- bzw. Hauptschule überwiesen).

c) Ein möglichst genaues Meßinstrument der Fremdsprachenkompetenz, das auch normorientierte Angaben machen kann, ist notwendig zur Differenzierung in Leistungsgruppen. Eine Zuweisung der Schüler nach den vom jeweiligen Lehrer abhängigen subjektiven Kriterien (die jeweils besseren Schüler in die A-Kurse, die jeweils schlechteren in die B-Kurse) hat häufig unerwünschte psychologische Auswirkungen und trägt der tatsächlichen fremdsprachlichen Kompetenz der Schüler in den wenigsten Fällen Rechnung. Auch hier ist zu untersuchen, ob der CT Lösungsmöglichkeiten bieten kann.

d) Traditionelle Meßverfahren sind häufig subjektiv oder aber, wenn sie objektiviert sind, nicht sehr aufschlußreich für eine Sprachkompetenz, die als Konstrukt integrierter Skills gesehen wird (vgl. 2.2.). Das Angebot an kommerziell erhältlichen Tests ist nicht sehr groß (z.B. Hamburger Englischtest, HET 6, 1966; Englisch Leistungstest 6 - 7, Beltz; Englisch Leistungstest 12 - 13, Beltz) und oft lehrwerkgebunden. Der Bereich der Mittelstufe (ca. 4 Jahre Englischunterricht) ist überhaupt nicht abgedeckt. Keiner der vorhendenen kommerziellen Tests ist speziell nach dem neuen Leitziel des Fremdsprachenunterrichts, der Kommunikativen Kompetenz, ausgerichtet [4b].

[4a] Vergleiche Hymes (in: Pride, 1974[2] : 28)

[4b] Vorhandenes Angebot siehe: "Leistungsmessung und Beurteilung im kommunikativen Englischunterricht". Hessisches Institut für Lehrerfortbildung. Reinhardswaldschule, 1979.

Roeder sagt dazu in seinem Gutachten für die Bildungskommission,

> "daß auch bei den besten gegenwärtig verfügbaren Tests das Fehlerrisiko zu groß ist, um individuelle Laufbahnentscheidungen auf sie zu gründen." (1974:109)

Seitdem hat sich wenig daran geändert.

1.1.2. Forschungsfragen

Es ergeben sich folgende Fragestellungen für die vorliegende Arbeit:

1) Ist der CT in unserem Schulsystem bei Schülern mit Deutsch als Muttersprache auch in Anfangsklassen des Englischunterrichts brauchbar?
2) Kann der CT eventuell an Stelle traditioneller Testmethoden, die jede 'Skill' einzeln messen, als Meßinstrument der Fremdsprachenkompetenz eingesetzt werden?
3) Entspricht der Test der Neuorientierung des Englischunterrichts im Hinblick auf das Leitziel der Kommunikativen Kompetenz?
4) Kann der CT zur Zuweisung von Schülern sowohl in verschiedene Schularten als auch in verschiedene Leistungskurse verwendet werden?
5) Könnte der CT dem Lehrer auch im Sinne eines informellen Tests zur Beurteilung der Fremdsprachenkompetenz dienen?
6) Was sind Vor- bzw. Nachteile des CT?
7) Was ist das beste Format des CT?

1.2. Definition und Umfang des Forschungsgegenstandes

1.2.1. Definition des Begriffes 'Cloze Test'

Obwohl die meisten Untersuchungen den CT W.L. Taylor (1953) zuschreiben, ist das Konzept älter. Einmal geht es auf die Gestaltpsychologie zurück, zum anderen kann man ein ähnliches Verfahren schon bei Ebbinghaus (1897) in seiner "Kombinationsmethode" finden. Es waren aber weder Taylor noch Ebbinghaus an 'Cloze' an sich interessiert. Sie benutzten das Verfahren zu anderen Zwecken. Taylor entwickelte daraus ein

Meßinstrument, um die Schwierigkeit von Prosatexten im muttersprachlichen Bereich zu bestimmen. Ebbinghaus wollte Unterschiede in der Konzentrationsfähigkeit von Schülern damit messen.

Das Wort cloze kommt von der Theorie der 'closure', die Taylor (1953:415) beschreibt als

> "the notion that humans tend to perceive a familiar pattern as a whole even when parts of it are missing, obscured or distorted".

Osgood (1965)[4c] definiert das Phänomen als "the tendency to fill in a gap in a well-structured whole". Auf dieser Grundlage wurde der CT entwickelt, den Taylor beschreibt als

> "a method of intercepting a message from a 'transmitter', mutilating its language patterns by deleting parts and so administering it to 'receivers' that their attempts to make the patterns whole again yield a considerable number of cloze units". (Taylor, 1953:416)

Oller (1975:2) beschreibt das Wesentliche des CT-Verfahrens konkreter:

> "It requires a subject or a group of subjects to replace missing verbal material in some context".

Diese Beschreibung dient der vorliegenden Arbeit als Grundlage, was die Verfahrensweise betrifft. Es soll an gegebener Stelle genauer erläutert werden, wie das Testverfahren gehandhabt wird.(Siehe 2.4.1.)

1.2.2. Umfang des Forschungsgegenstandes und Eingrenzung

Das Cloze-Verfahren hat einen sehr breiten Anwendungs- und Forschungsbereich gefunden, der wie folgt gegliedert werden kann:

TAFEL 1

I Allgemeine Forschung und Diskussion der Grundproblematik

II Cloze im Englischunterricht als
- Meßinstrument der Fremdsprachenkompetenz
- Lehrinstrument
- Meßinstrument zur Feststellung der Textschwierigkeit (Readability)
- Instrument, bestimmte Reaktionen hervorzurufen (Elicitation device)
- Meßinstrument im Hör-Verstehensbereich

III Cloze im muttersprachlichen Bereich

IV Cloze im Bereich anderer Fremdsprachen und verschiedener Dialekte und Sozialformen der Sprache

[4c] nach James, 1979:1

Dem oben gesteckten Ziel der Arbeit entsprechend bezieht die vorliegende Dissertation nur die Anwendungsbereiche unter 1, soweit relevant, und den CT als Meßinstrument der Fremdsprachenkompetenz ein. Die verbleibenden Anwendungsbereiche werden hier bewußt ausgeklammert, denn zu ihrer Bearbeitung bedarf es eines besonderen Versuchsinstrumentariums und einer anders gerichteten Versuchsdurchführung. Cloze-Verfahren sind sowohl im graphischen wie im phonetischen Bereich der Sprache durchführbar. Zur Reduktion der Variablenmenge und der besseren Durchführbarkeit und Verfügbarkeit wegen wird hier der CT nur im Bereich der geschriebenen Sprache verwendet und untersucht.

1.2.3. Standardmodell des CT

Als Meßinstrument zur Bestimmung der Fremdsprachenkompetenz erscheint der CT in einer Vielzahl von Varianten. Die folgende Aufstellung dient als kontrastive Darstellung der 'differentia specifica' und zur Klarstellung der Terminologie.

TAFEL 2

Art des Kontextes	Art des 'Cloze Item'
1 Isoliertes Wort oder Lexem	Auslassen eines beliebigen Teiles
2 Beschränkter Kontext (z.B. Satz, Äußerung)	Auslassen eines einzelnen Wortes oder beliebiger anderer Einheiten (Lexem, Morphem, etc)
3 Mehrsätziger Text mit fortlaufendem Inhalt	a) Auslassen jedes n-ten (normalerweise jedes 5., 6. oder 7.) Wortes (Taylor, 1953)
	b) Auslassen spezifischer Wortklassen
	c) Auslassen eines Wortes pro n-ten Satz (z.B. Clausing und Lange, 1977)
	d) Auslassen bestimmter Wörter, die den Trennschärfekriterien der Testtheorie entsprechen (Evans und Haastrup, 1976)
	e) Auslassen von Wörtern oder anderen Elementen nach Belieben (Ebbinghaus, 1897)
	f) zusätzliches Einsetzen von Elementen

in Wörter

g) Ersetzen einzelner Elemente durch anomale Elemente

h) zusätzliches Einsetzen anomaler Elemente

Oller (1979) bezeichnet das Verfahren, das hier unter 3a eingeordnet ist, als Standardmodell, das in den meisten Studien verwendet wurde. Es ist das in dieser Studie verwendete Verfahren.

Der Diskussion methodischer Detailfragen und der Frage nach der Operationsweise des CT ist ein eigener Abschnitt in Kapitel 2 eingeräumt.

1.3. Bearbeitungsschritte - Forschungsmethoden

Die Bearbeitung des umrissenen Themas erfordert folgende Schritte: Einen Überblick über die Literatur verschiedener relevanter Fachgebiete. Im Vordergrund steht zunächst die Frage, wozu Fremdsprachentests gebraucht werden. Die Frage nach der Brauchbarkeit setzt eine Bestimmung der Gütekriterien von Tests voraus, die aus der psychometrischen Theorie entnommen werden. Die Fremdsprachentesttheorie bestimmt die Art und Weise, wie getestet werden kann.

Die Frage, ob der CT die Fremdsprachenkompetenz entsprechend den Leitzielen des Unterrichts erfassen kann, erfordert eine Bestimmung des Kompetenzbegriffs.

Die Frage nach der Verwendbarkeit des CT in unserem Schulsystem erfordert eine Bestimmung der wesentlichen relevanten Merkmale des Schulsystems.

Eine Untersuchung des Forschungsgegenstandes kann erst erfolgen, wenn seine methodischen und operationalen Charakteristika dargestellt sind, wie sie in der Literatur und Forschung vorzufinden sind.

Auf der Grundlage dieser Ausführungen kann die Problematik erfaßt und können Hypothesen erstellt werden. Eine Versuchsanordnung wird entworfen und durchgeführt, um die Gültigkeit der Hypothesen zu überprüfen. Die Resultate dieses Versuchs sollen einen Rückschluß auf die gestellten

Forschungsfragen erlauben und Hinweise auf Möglichkeiten und Notwendigkeiten weiterer Forschung geben.

1.4. Umfang der Studie

1.4.1. Material

Das Sammeln des Datenmaterials erstreckte sich auf die Jahre 1977 bis 1980. Cirka 16 Tests wurden entworfen. Sie wurden an drei verschiedenen Hauptschulen, einer Realschule, drei Gymnasien und der Universität München (Einstufungstestkandidaten) durchgeführt. Dabei wurden insgesamt ca. 80 Testdurchgänge in ungefähr 30 Klassen mit 800 Schülern durchgeführt. Es wurden 2000 Testbogen mit insgesamt 50 000 Einzelitems gesammelt.

1.4.2. Eingrenzung der Studie

Da, wie eingangs erwähnt, völliges Neuland betreten wurde, erfolgte ein Teil der experimentellen Arbeit nur aufgrund informeller Hypothesen in explorativem Stil. Die Erfahrung daraus zeigte, daß das ursprüngliche Vorhaben, den CT auf allen Stufen unseres Schulsystems zu erproben, zeitlich und organisatorisch durch eine Einzelperson nicht möglich gewesen wäre. Vor allem hätte die Genauigkeit sowohl der Durchführung als auch der Auswertung gelitten. Bei der Vielzahl der Variablen wären außerdem praktikable Ergebnisse kaum zu erwarten gewesen und die Übersicht völlig verloren gegangen. Aus diesen Gründen erwies es sich als notwendig, sich auf die Klassen 5 - 9 unseres dreigliedrigen Schulsystems zu konzentrieren, die wie folgt verteilt sind:

	5. Jahrgang	6.J.g.	7. J.g.	8. J.g.	9. J.g.
Hauptschule	×	×	×	×	×
Realschule			×	×	×
Gymnasium	×	×	×	×	×

1.4.3. Weiterentwicklung der Studie

Trotz des Umfangs des gesammelten Materials ist die vorliegende Arbeit eine Pilotstudie. Sie soll die Grundlagen schaffen für eine Weiterentwicklung anwendungsreifer standardisierter CTs.

B e s o n d e r e A n m e r k u n g e n :

Die Vielzahl der Termini technici, die größtenteils aus der amerikanischen Forschung stammen, wo sie fest etabliert sind, haben häufig keine deutschen Entsprechungen. In Fällen, wo deutsche Übersetzungen erfolgt sind, hat dies teilweise zu einem Verlust an der Anschaulichkeit und Präzision geführt, oder es sind auch miteinander konkurrierende Begriffe eingeführt worden (siehe z.B. im Kapitel zur Intentionalität). Es erscheint deshalb zweckmäßig und gerechtfertigt, zur Vermeidung weiterer Unklarheiten die Originalbegriffe beizubehalten, wenngleich dies dem ästhetischen Betrachter als Stilfehler erscheinen mag. Viele der Begriffe sind außerdem mittlerweile in den deutschen Fachwortschatz integriert und dürften kaum Verständnisprobleme auslösen.

2. VORGEFUNDENE VORAUSSETZUNGEN IN DER LITERATUR UND FORSCHUNG

Ist es das allgemeine Ziel dieser Arbeit, den CT als Meßinstrument der Fremdsprachenkompetenz von Schülern unseres Schulsystems zu erforschen, so ist es die Aufgabe dieses Teiles, die Grundproblematik dieses Arbeitsfeldes darzustellen, wie sie in der Literatur vorzufinden und was für das Projekt daraus abzuleiten ist.
Auf dieser Basis können Hypothesen erstellt und experimentell überprüft werden. Dies darzustellen ist die Aufgabe der darauffolgenden Teile.

2.1. Grundlagen des Fremdsprachentestens: Intentionalität - Qualität - Modalität

2.1.1. Intentionalität

Da Fremdsprachentests keinen intrinsischen Wert besitzen, sondern ihren Wert erst aus ihrem Nutzen ableiten, ist ein erster übergeordneter Betrachtungspunkt die Testintentionalität.
Verschiedene Autoren geben hier unterschiedliche Antworten und Klassifizierungsmodelle, wobei die Terminologie nicht einheitlich ist. Die folgende Übersicht präsentiert die Hauptvertreter der Fremdsprachentesttheorie und vergleicht ihre Modelle:

HARRIS diskutiert sechs Hauptziele von Sprachtests:

"1. To determine readiness for instructional programs.
2. To classify or place individuals in appropriate language classes.
3. To diagnose the individuals' specific strengths and weaknesses.
4. To measure aptitude for learning.
5. To measure the extent of student achievement of the instructional goals.
6. To evaluate the effectiveness of instruction." (1969:2)

Zur Vereinfachung ordnet er diese sechs Kategorien in drei Gruppen:
1) Aptitude Tests (siehe oben 4.)
2) General Proficiency Tests (siehe oben 1.-3.)
Es ist bemerkenswert, daß nach Harris diese Testart auch ein prognostisches Element besitzt, denn sie kann als Basis dienen "for future attain-

ment" . Die dritte Gruppe sind nach Harris die Achievement Tests (siehe oben 5 und 6).

VALETTE (1967, 1977) unterscheidet vier Arten von Sprachtests:
1) Aptitude Tests (1967 auch Prognostic Tests). Im Gegensatz zu Harris, bei dem Aptitude Tests auch die Funktion haben, zukünftige Sprachleistung zu messen, sollen sie bei Valette nur die Fremdsprachenlernfähigkeit messen. Naturgemäß ergeben sich durch das Hinzufügen eines vierten Tests Verschiebungen im übrigen Bereich. Der Begriff "General Proficiency Test" von Harris divergiert bei Valette in Proficiency Test mit einer Subkategorie Placement Test (entspricht aber dem Punkt 2 bei Harris, siehe oben) und Achievement Test. Eine besondere Komponente des Proficiency Test bei Valette ist seine Gerichtetheit auf bestimmte, vom Probanden in Zukunft zu erfüllende Aufgaben, z.B. ob er in der Lage ist, Vorlesungen in der Fremdsprache zu folgen. Im Deutschen wird hierfür von Brings (1970) der Terminus Qualifikationstest verwendet. W. Kahl (Übersetzung von Valette, 1967, erschienen auf deutsch 1971) übersetzt den Begriff mit Spezifischer Leistungstest. Valettes Proficiency Test[4d] stuft den Probanden außerdem nach funktionalen Gesichtspunkten ein, z.B. wie er in der Fremdsprache funktioniert - ein vages, aber doch bemerkenswertes Kriterium, da es bereits pragmatische Elemente enthält.
Die Definition von Harris für Achievement Test (siehe oben 5. und 6.) trifft zu auf den Progress Test von Valette mit der Zusatzkomponente, daß er vom Klassenlehrer entworfen ist, während Valettes Achievement Test eine Sondervariante des Progress Test ist, mit der Komponente, daß er für Probanden verschiedener Herkünfte und Hintergründe, also nicht kursorientiert ist. Der diagnostische Zweck des Proficiency Test bei Harris (siehe oben 3) und der oben erwähnte prognostische Aspekt werden bei Valettes Proficiency Test nicht erwähnt.

HEATONS (1975) Definition des Proficiency Tests deckt sich weitgehend mit der von Valette, besitzt aber die zusätzlichen Komponenten, daß er nicht kursorientiert ist und für Lernende verschiedener Sprachlehrhintergründe[3a] entworfen ist, welche letztere Komponente aber Valettes Achievement Tests besitzen.

[4d] Hier bezogen auf das Beispiel des Tests des Foreign Service Institute

Progress Tests bei Heaton decken sich mit Progress Tests bei Valette, besitzen aber zusätzlich die Aufgabe, zu diagnostizieren und den Schüler zu motivieren. Heatons Achievement Test (zusätzlicher Terminus: Attainment Test), deckt sich mit dem von Harris, hat aber die zusätzlichen Komponenten größerer Formalität und nicht vom Klassenlehrere entworfen zu sein. Darüberhinaus konstatiert Heaton, daß viele Progress -, Achievement - und Proficiency Tests auch als diagnostische Tests verwendet werden können.

DAVIES' (1977) Proficiency Test deckt sich mit dem Heatons, der Achievement Test (oder Attainment Test) mit Punkt 5 von Harris. Davies weist keinen besonderen Progress Test aus, etabliert aber dafür Diagnostic Tests als eigene Testart. Sie sind definiert als eine Art "negativer achievement tests", die nach Fehlerlisten aus Achievement und Proficiency Tests entworfen sind und Strukturen von Fehlleistungen überprüfen sollen.

Die Beurteilung der Fähigkeit, wie ein Proband die Sprache in "real-life" Situationen benützen kann, ist die Aufgabe der Proficiency Tests bei CLARK (1978). Er unterteilt dabei die Tests in zwei Verfahren

a) direkte Verfahren, die "real-life" Situationen so ähnlich wie möglich sein sollen, und

b) indirekte Verfahren, die nicht unbedingt identisch sein müssen.

Zu diesen Verfahren zählt er expressis verbis den CT. Er begründet dies damit, daß das Cloze Verfahren nur in den seltensten Fällen dem tatsächlichen Sprachgebrauch entspricht, d.h. daß sehr selten in einem gehörten oder gelesenen Text durch verschiedene Störfaktoren Lücken entstehen. Das mag zutreffen, aber es kann genauso argumentiert werden, daß die Situation für den Sprachlernenden so aussieht: Jedes unbekannte Wort stellt sozusagen eine Lücke dar, für die er Lösungsstrategien entwickeln muß. Ein weiteres Argument dagegen, den CT als nicht lebensnahe Testform zu betrachten, bringt Oller (1979) mit der sogenannten "Expectancy Grammar", die er als "real-life" Gebrauch der Sprache diskutiert,

allerdings auf einer Nicht-Oberflächenebene (siehe 2.1.3.4.).
Clarks Achievement Test entspricht dem Punkt 5 (siehe oben) von Harris, wobei er aber zwei Subklassifizierungen hinzufügt:

a) den Diagnostic Achievement Test, der Lücken im Test diagnostiziert, und
b) den General Achievement Test, der die Sprachfähigkeit des Schülers unter dem Aspekt des Kursinhalts

mißt, aber dabei dem tatsächlichen Sprachgebrauch auf eine realistischere und globalere Art und Weise näherkommt.

LADO (1961) weist drei Hauptgruppen von Sprachtests aus: Achievement Tests, Diagnostic Tests und Aptitude Tests. Seine Definition des letzteren Tests beinhaltet sowohl den Faktor der Sprachlernfähigkeit als auch den der prognostizierten Leistung. Die beiden anderen Tests stehen aber nebst Subklassifizierungen in besonderer Beziehung zueinander. Beide sind zusammengefaßt unter dem Oberbegriff Achievement Tests[5].

> "Making reference to the fact that students have to struggle through a course or a learning experience of some sort to 'achieve' a certain amount of control of the language. When the same tests are thought of independently of the learning experience, they may then be referred to as proficiency test[5a]. Proficiency tests measure how much of a foreign language a person (not necessarily a student) knows." (Lado, 1961:369)

Eine Unterart bildet der Placement Test (übersetzt von Freudenstein als Einstufungs- oder Eingangstest, welcher aber trotzdem noch je nachdem ein Achievement oder Proficiency Test ist. Diagnostic Tests[5b] (übersetzt von Freudenstein, 1971) sind ebenso Achievement Tests mit der zusätzlichen Komponente, besondere Stärken und Schwächen innerhalb der gemessenen Fertigkeit nachzuweisen. Diese Stärken und Schwächen können sehr spezifisch sein, wie z.B. die Beherrschung einzelner Phoneme, oder sehr allgemein, wie z.B. Lese- oder Schreibfähigkeit in der Fremdsprache.

Das Dargestellte läßt die verwirrende terminologische Vielfalt erkennen, die durch die Übertragung ins Deutsche keinesfalls vereinfacht wurde. Die folgende Matrix stellt einen Versuch dar, Übersichtlichkeit zu schaffen.

---

5 übersetzt von Freudenstein, 1971, als 'Leistungstest'

5a übersetzt von Freudenstein als 'allgemeiner Leistungs- oder Abschlußtest'

5b übersetzt von Freudenstein als 'Diagnostiktest'

MATRIX I

Testart \ Autor	HARRIS (1969)	VALETTE (1967/77)	HEATON (1975)	DAVIES (1977)	CLARK (1978)	LADO (1961)
APTITUDE/PROGNOSTIC TESTS nach Brings: EIGNUNGSTEST	Bezeichnung: APTITUDE TEST screening test/ to predict future performance/ to assess an individual's (little or no knowledge) performance in the future	Bezeichnung: PROGNOSTIC OR APTITUDE TEST Übersetzt von Kahl: Eignungstest to select students/ to measure language learning ability (little or no knowledge)	Bezeichnung: APTITUDE TEST measures student's probable performance/ shows aptitude	Bezeichnung: APTITUDE TEST no syllabus at all tests construct of language aptitude	Bezeichnung: APTITUDE TEST prognostic function to predict level of likely accompilshmert	Bezeichnung: APTITUDE TEST PROGNOSTIC TEST Übersetzt von Freudenstein: EIGNUNGSTEST to predict future success through measuring individuals' learning ability
PROFICIENCY TESTS nach Brings: QUALIFIKATIONSTEST Subkategorie EINSTUFUNGSTEST	GENERAL PROFICIENCY TEST 1) determine readyness for instruction 2) classify or place individuals in appropriate lang. classes 3) to diagnose the individual's specific strengths and weaknesses serves as basis for predicting future attainment	PROFICIENCY TEST Übersetzt von Kahl: SPEZIFISCHER LEISTUNGSTEST measures what student has learned; aim: determine if language ability corresponds to specific requirements subcategory: PLACEMENT TEST Übersetzt von Kahl: EINSTUFUNGSTEST usually reflects student's ability on a certain continuum that reflects a predetermined set of categories measures the functioning of student in the foreign language (67: should be designed by school or college itself)	PROFICIENCY TEST measures student's achievement in relation to specific task he will later be required to perform not syllabusoriented often administered to students from various language learning backgrounds	PROFICIENCY TEST to establish certain standards for specific task later to perform not syllabus oriented must be available to all students from different backgrounds	PROFICIENCY I to determine the student's ability to use the language for real-life purposes focuses on examinee's ability to perform pragmatically useful tasks in the language a) direct procedures approximate real-life situation in four modalities b) indirect procedures do not require authentical use e.g. cloze Evaluation of acquisition of language based on previous formal instruction but procedures applicable to any course involves sampling subcategories: a) DIAGNOSTIC ACHIEVEMENT TEST measures the lack of acquisisition of discrete items b) GENERAL ACHIEVEMENT TEST to measure student s ability to combine aspects of content in situations which more closely approximate ordinary language use more global and realistic	PROFICIENCY TEST Übersetzt von Freudenstein: ALLGEMEINER LEISTUNGS- ODER ABSCHLUSSTEST Independent of learning experience measures achievement of amount of labguage control of a person open for any person, not necessarily a student subcategory: PLACEMENT TEST Übersetzt von Freudenstein: EINGANGSTEST
ACHIEVEMENT ATTAINMENT	ACHIEVEMENT TEST 5) measures extent of achievement of instructional goals 6) evaluates effectiveness of instruction	ACHIEVEMENT TEST ÜBERSETZT von Kahl: ALLGEMEINER LEISTUNGSTEST similar to progress ofttest measures the achievement not of specific teaching materials but for students with different backgrounds	ACHIEVEMENT OR ATTAINMENT TEST to show mastery of specific syllabus (end-of-year, public tests) similar to progress test but not designed by classroom teachers more formal	ACHIEVEMENT OR ATTAINMENT TEST samples items from the content of specific syllabus of course (end-of-year)		ACHIEVEMENT TEST auch übersetzt als: LEISTUNGSTEST Is in reference to some course content, measures achievement of certain amount of language control subcategory: PLACEMENT PROFICIENCY TESTS are also achievement tests but independent of learning experience
PROGRESS TEST	see: DIAGNOSTIC ACHIEVEMENT TEST	PROGRESS TEST Übersetzt von Kahl: LERNFORTSCHRITTSTEST how much student has learned in specific course of instruction designed by classroom teacher	PROGRESS TEST to assess progress of students in class. most classroom tests are progress tests to evaluate success of teaching identify weakness or difficulty given to motivate student	see: DIAGNOSTIC ACHIEVEMENT TEST		
DIAGNOSTIC TEST	see 3 above: PROFICIENCY TEST	see: PROGRESS TEST	DIAGNOSTIC TEST many progress, achievement and proficiency tests can be used as diagnostic tests to identify areas of weakness or difficulty/ used to plan instruction primarily to assess student's knowledge of skills before a course is begun, species of "nonachievement test" based on list of errors made in achievement or proficiency test, consists of multiple items of frequent errors to check patterns of errors		see: DIAGNOSTIC ACHIEVEMENT TEST	DIAGNOSTIC TESTS are also achievement tests, characterized by the distinctive feature "to show specific weaknesses or strengths"

Die aufgezeigte Unterschiedlichkeit der Termini läßt erkennen, daß sie für die weitere Forschungsarbeit kaum brauchbar und nicht ohne weiteres auf das Forschungsgebiet anwendbar sind. Dabei tragen die verschiedenen deutschen Übersetzungen eher zur Verwirrung als zur Klärung bei. Insbesondere fließen in die Definitionen der verschiedenen Autoren in unkontrollierter Art Elemente aus verschiedenen Bereichen der Intentionalität ein.

In der folgenden Matrix[2], die weiterhin als Arbeitsgrundlage dienen soll, werden deshalb die einzelnen Bereiche klar getrennt, und die entsprechenden Elemente jeweils zugeordnet. So können terminologische Schwierigkeiten vermieden werden, Einzelfaktoren, je nach Bedarf, klar dargestellt und in übersichtlicher Weise miteinander kombiniert und weiterhin spezifiziert werden.

Eine 'Faktorenanalyse' läßt vier Grundbereiche erkennen:

1) Testfunktion (functions)
2) Zweck (uses)
3) Zielgruppe (target group)
4) Form (formality)

TEILBEREICHE DER INTENTIONALITÄT VON FREMDSPRACHENTESTS
(MATRIX 2)

FUNKTION	ZWECK	ZIELGRUPPE	FORM
Der Test dient als direktes oder indirektes Meßinstrument a) der Fremdsprachenlernfähigkeit b) zukünftiger Fremdsprachenkompetenz c) allgemeiner Fremdsprachenkompetenz d) der Beherrschung einzelner Teile der Fremdsprachenkompetenz Der Test dient als Motivations- und Lehrinstrument	a) Eignungsvorhersage b) Diagnose von Stärken und Schwächen des Lernenden c) Kontrolle von Lernzielen d) Kontrolle der Unterrichtseffektivität e) Einstufung von Individuen für bestimmte Zwecke f) Einstufung von Gruppen für bestimmte Zwecke g) Schaffung und Vergrößerung der Lernmotivation	a) Probanden mit gleichem fremdsprachlichen Lehr- und Lernhintergrund b) Probanden mit unterschiedlichem fremdsprachlichen Lehr- und Lernhintergrund c) Probanden ohne fremdsprachlichen Lehr- und Lernhintergrund (siehe 1a)	a) vom Lehrer zusammengestellte Aufgabensammlung b) informelle Tests c) standardisierte Tests

Beim Vergleich von Matrix 2 mit Matrix 1 sind jetzt klar Lücken und Unterschiede in den Definitionen der einzelnen Autoren zu erkennen. So wird zum Beispiel die als Aptitude Test (gelegentlich auch als Prognostic Test bezeichnet) klassifizierte Testart von Valette und Davies nur durch die Funktion a der Matrix 2 bezeichnet, von Heaton und Clark nur durch Funktion b und von Harris und Lado von a und b. Es ist auch erkennbar, daß die Begriffe Aptitude Test und Prognostic Test nicht ohne weiteres ausgetauscht werden können, wie dies geschieht. Funktion a bezieht sich eher auf den Begriff Aptitude Test, Funktion b auf Prognostic Test. Deshalb ist auch die Konvergenz beider Begriffe im Deutschen zu Eignungstests nicht ganz zufriedenstellend. Wenngleich jeweils beide Tests den gleichen Zweck verfolgen, nämlich die Eignungsvorhersage und die Einstufung für bestimmte Zwecke, muß das Testinstrument aber verschieden konstruiert sein, um den beiden Funktionen jeweils dienen zu können. So kann die künftige Fremdsprachenkompetenz nur indirekt gemessen werden. Dies geschieht etwa, indem Korrelationen zwischen (meßbaren) gegenwärtigen sprachlichen Verhaltensweisen, gegebenenfalls auch im muttersprachlichen Verhalten und vorauszusagenden fremdsprachlichen Verhaltensweisen nachgewiesen werden. Eine vorhandene Fremdsprachenlernfähigkeit kann auch direkt gemessen werden, etwa durch die Bestimmung grundlegender Sprachlernstrategien (Überblick bei Hennon et al, 1929; Salomon, 1954; Pimsleur et al, 1962). Wird zum Beispiel das Cloze Verfahren als solches erkannt, könnte es auch auf diesem Gebiet eingesetzt werden. Momentan liegen keine derartigen Untersuchungen vor. Auch im Rahmen dieses vorliegenden Projekts sind sie nicht vorgesehen. Ein Hinweis auf Forschungsmöglichkeiten sei aber gemacht.[6]

Obwohl die Funktion eines Tests als Motivationsinstrument nur selten aufgeführt ist (z.B. Heaton, 1975), wird sie bewußt in die Matrix einbezogen. Es ist gerade dieses Element, das die Verbindung zu dem pädagogischen Rahmen, in dem er eingesetzt wird, herstellt, und das von beson-

---

6 Vergleiche zum Gesamtkomplex der Fremdsprachenbegabung u.a. Carroll, 1973; Pimsleur, 1976; Jakobovits, 1971; Valette, 1977 und Scheibner-Herzog/Pieper, 1977.

derer Wichtigkeit ist, wenn man die Sprache nicht nur als Meßobjekt sondern auch in ihrem Wesen als soziales Phänomen betrachtet.
Gleichzeitig ist hier auf einen besonderen Aspekt des Motivationsfaktors hinzuweisen; der sogenannte 'backwash' (gelegentlich auch 'washback'). Damit wird der Einfluß bezeichnet, den der Test auf das Lehren und Lernen, das vor dem Test stattfindet, bezeichnet. Dieser Einfluß kann sowohl negativ als auch positiv sein. Das erstere trifft zu, wenn vom Ziel des Unterrichts abgelenkt wird, und der Test ein eigenes unabhängiges, eventuell sogar unerwünschtes Lernziel bildet, z.B. das traditionelle Diktat, oder dann, wenn der Test nur in der Lage ist, eine instrumentelle oder Sekundärmotivation zu schaffen; d.h. wenn nicht der Textinhalt, sondern der durch den Test geschaffene Druck motivierend wirkt.
Der Backwasheffekt ist dann positiv, wenn das Testziel mit dem Unterrichtsziel übereinstimmt und der Test die Schüler zusätzlich motiviert.
In einem kybernetischen Unterrichtsmodell bildet der Test durch seine Möglichkeit, ein 'feedback' zu liefern, ein wesentliches Element des Regelkreises. Ein Lehr- und Lerninstrument im Sinne des Wortes bildet ein Test dann, wenn er es durch seine Problemstellung vermag, dem Lernenden beim Lösungsvollzug zu einem Lernzuwachs zu verhelfen. Vom CT wird allgemein angenommen, daß er einen positiven Backwasheffekt hat.

Trotz einer formalen Aufgliederung sind die Elemente der Matrix 2 überlappend und miteinander verbunden und auch gegenseitig abhängig.
So ist zu fragen, ob die Prüfung einer allgemeinen Fremdsprachenkompetenz auch die Messung einzelner Teile der Fremdsprachenkompetenz erfordert. Dies erfordert eine Bestimmung des Konstruktes der Fremdsprachenkompetenz, woraus sich wiederum die Durchführungsmodi ableiten lassen.
Eine nähere Diskussion von Zweck und Zielgruppe kann nur im Blickpunkt auf schulpädagogische und fachdidaktische Standorte gesehen werden.
Die Testqualität steht in funktionaler Abhängigkeit von ihrer Formalität. Diese Diskussion wird in den nächsten Abschnitten durchgeführt.

Ein idealer Test sollte mit größter Genauigkeit und geringstem Aufwand das messen, was er zu messen beabsichtigt. Aus dieser Feststellung ist zu entnehmen, daß die Qualität eines Tests durch mehrere Kriterien bestimmt ist.

2.1.2. Qualitätskriterien von Tests

2.1.2.1. Allgemeine Übersicht

Die Testintentionalität mit ihren oben aufgeführten Teilbereichen bestimmt die Höhe der Qualitätsanforderungen an einen Test. So sind die Anforderungen an eine vom Lehrer zusammengestellte Aufgabensammlung, die einen engen Teil der Fremdsprachenkompetenz, etwa einen gewissen Wortschatz, zur Lernzielkontrolle überprüft, nicht so hoch wie an einen Einstufungstest, der die weitere berufliche Zukunft des Schülers bestimmt.

Die Hauptvertreter der Fremdsprachentesttheorie weisen auch hier unterschiedliche Kriterien aus.

LADO (1961) weist fünf Kriterien für die Bewertung von Sprachtests aus:

1) Validity — Gültigkeit (übersetzt von Freudenstein)
2) Reliability — Zuverlässigkeit
3) Scorability — Auswertbarkeit
4) Economy — ökonomische Vertretbarkeit
5) Administrability — organisatorisch-technische Durchführbarkeit

Nach HARRIS konstituieren drei primäre Kriterien das sine qua non von Tests:

1) Validity
2) Reliability
3) Practicality [7a]

VALETTE (1977) nennt zwei wesentliche Charakteristika eines guten Tests: [7b]

1) Reliability → (übersetzt von Kahl als Zuverlässigkeit oder Reliabilität)

   Der Test muß verläßliche Werte ergeben.

2) Validity → (Gültigkeit oder Validität)

   Der Test muß wirklich das messen, was er messen soll.

HEATON führt vier Charakteristika eines guten Tests auf:

1) Validity
2) Reliability
3) Discrimination
4) Backwash

HEATON (1975)

7a Im Punkt 3 wird hier die gerade im schulpraktischen Bereich hilfreiche Differenzierung Lados durch die Zusammenfassung unter dem Begriff der Practicability zu ungunsten der Vergröberung aufgegeben.

7b Valette bezieht sich auf Henry Chauncey und John E. Dobbin, Testing. Its Place in Education Today. (New York, 1963)

Mit 'Discrimination' bezeichnet Heaton die Fähigkeit eines Tests, daß im Gegensatz zu einem Classroomtest, wo es primär darauf ankommt, zu zeigen, ob eine Klasse ein Ziel insgesamt erreicht hat oder nicht, einzelne Prüflinge in ihrer Leistung voneinander unterschieden werden (vgl. Heaton, 1975: Für den Bereich unseres Schulsystems kann diese Beschreibung nicht ohne weiteres übernommen werden, denn hier wünscht der Lehrer häufig auch bei Classroomtests eine gewisse Trennschärfe, nicht zuletzt wegen der großen Bedeutung dieser Tests für die Laufbahn der Schüler.

DAVIES (1977) nennt

Validity und
Reliability

unter der Überschrift "Test Virtues".

Validität und Reliabilität sind bei allen Autoren aufgeführte Kriterien. Unterschiede, zumindest oberflächlicher Art, bestenen bei den übrigen Kriterien. Bei genauerer Betrachtung ist aber zu erkennen, daß alle Kriterien - mit Ausnahme des bei Heaton aufgeführten 'backwash', welcher eher dem pädagogisch-didaktischen Bereich zuzuordnen ist (siehe 2.1.1.) - der psychometrischen Testtheorie entstammen.[8] Auf sie beziehen sich alle Autoren und wählen Wesentliches für die spezifischen Bedürfnisse des Fremdsprachentestens aus. Da im Rahmen dieser Arbeit eine detailliertere Diskussion der psychometrischen Testtheorie erforderlich ist, bietet es sich an, die dort vorliegenden Ergebnisse als Grundlage für die weitere Arbeit zu verwenden.
Im deutschen Sprachbereich ist vor allem die Terminologie von Lienert (^{2}1967) gebrauchlich. Sie basiert auf einer umfassenden und präzisen Darstellung von Testaufbau und Testanalyse. Bei Lienert liegt auch eine brauchbare Definition von 'Test' vor:

> "Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung." (Lienert, 1967:7)

Nach Lienert (1967:8) soll ein guter Test folgende drei Forderungen als Hauptgütekriterien erfüllen; er soll

1) objektiv
2) reliabel
3) valide sein.

---

8 Standardwerke der psychometrischen Testtheorie sind besonders: Henry Chauncey/John E. Dobbin, Testing, Its Place in Education Today, (New York, ^{2}1964); Günter Clauss/Heinz Ebner, Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen (Frankfurt/Zürich, 1970); Gustav Lienert, Testaufbau und Testanalyse (^{2}1967); J.C. Nunnally, Psychometric Theory (New York, ^{2}1972).

Daran schließen sich vier "Nebengütekriterien" als bedingte Forderungen an: der Test soll

4) normierbar
5) vergleichbar
6) ökonomisch
7) nützlich sein.

Lienerts Einteilung der Kriterien unterscheidet sich - zumindest oberflächlich - von den oben aufgeführten Einteilungen. Insbesondere fällt die Einführung des Kriteriums der Objektivität und seine Plazierung an erster Stelle auf. Diese Ausweisung der Objektivität als separates Kriterium ist im deutschsprachigen Raum gebräuchlich.[9]

2.1.2.2. Objektivität

Das Gütekriterium der Objektivität bildet die Grundlage für die in der Testdefinition genannte Möglichkeit einer stabilen quantitativen Aussage über eine Merkmalsausprägung von Probanden.

Lienert definiert Objektivität

> "Unter Objektivität eines Testes verstehen wir den Grad, in dem die Ergebnisse eines Testes unabhängig vom Untersucher sind. (Lienert, [2]1967:13)

Je nach Phase der Testdurchführung werden verschiedene Aspekte der Objektivität unterschieden (Lienert, [2]1967:13,14):

1) die sog. Durchführungsobjektivität, die den Grad der Unabhängigkeit der Testergebnisse vom Untersucher während der Testdurchführung betrifft.

2) Die sog. Auswertungsobjektivität betrifft die numerische oder kategoriale Auswertung des registrierten Testverhaltens nach vorgegebenen Regeln. Bei Alternativentscheidungen (yes/no, true/false) ist die Auswertungsobjektivität praktisch vollkommen entwickelt. Etwas geringer ist sie bei Leistungstests mit freier Aufgaben- oder Fragebeantwortung.

3) Die sog. Interpretationsobjektivität betrifft den Grad der Unabhängigkeit der Interpretation des Testergebnisses von der interpretierenden Person, die nicht mit dem Untersucher

---

9 Siehe dazu Clauss/Ebner (1969:29,30); R. Dieterich (1972:52,84); Kamratowsky et al (1973:7); Ingenkamp (1973)

oder Auswerter identisch zu sein braucht.
Bei den eingangs aufgeführten Fremdsprachentestautoren ist das Kriterium der Objektivität nirgendwo explizit als solches erwähnt. Die einzelnen Aspekte erscheinen aber - meist implizit - unter anderen Gesichtspunkten.

Insbesondere wird, wie oben erwähnt, die Reliabilität in Abhängigkeit von Bedingungen der Objektivität gesehen. Auch Lienert (^{2}1967:211) weist darauf hin, daß die verschiedenen Arten der Objektivität als Sonderfälle der Reliabilität aufzufassen sind.

Bei VALETTE (1977:44) ist Objektivität, die aber nicht explizit genannt ist, eine Vorbedingung der Reliabilität. Sie unterscheidet vier Voraussetzungen für einen zuverlässigen Test:

1) multiple samples
2) standard tasks
3) standard conditions
4) standard scoring.

Punkt 1 bildet eine Sonderstellung: Obwohl die quantitative Beschaffenheit der Stichprobe vom Untersucher abhängig ist, hat dies wenig mit ihrer Qualität, bzw. mit der Art ihrer Reaktion, zu tun. Er ist deshalb in den Bereich der Statistik zu verweisen.

HEATON (1975:156) zählt folgende Faktoren auf, die im Sinne der Objektivität die Reliabilität eines Tests beeinflussen:

1) the extent of the sample
2) the administration of the test
3) test instructions
4) personal factors such as motivation and illness
5) scoring the test.

Faktor 4 ist ein endogener Faktor, der teilweise durch den Untersuchenden beeinflußt werden kann und deshalb eine Sonderstellung einnimmt.

HARRIS nennt zwei Faktoren aus dem Bereich der Objektivität, die die Reliabilität beeinflussen:

1) die tägliche Fluktuation z.B. der Testbedingungen oder Schülermotivation,
2) die Konstanz der Verhaltensweisen von Proband und Prüfer.

LADO fordert Objektivität im Dienste des von ihm aufgestellten Gütekriteriums der 'scorability', was bei Lienert der Auswertungsobjektivität entspricht. Bei Lado erscheint das Kriterium bei der Diskussion der Teststrategien: es gibt objektive und subjektive Tests (1961:28,29). Dabei erscheinen subjektive Fremdsprachentests oft natürlicher und oberflächlich

valider als objektive Tests. Objektive Tests werden oft äußerlich unnatürlich, da die Aufgaben auf mechanische Bewertungsformen zugeschnitten sind. Um trotz mangelnder äußerlicher Validität objektive Verfahren zu rechtfertigen, empfiehlt Lado, die linguistischen Probleme, die das Testziel bestimmen, linguistisch genau zu analysieren. Wenn sich dann eine Aufgabe als linguistisch valide erwiesen hat, kann das auch eine größere Objektivität trotz mangelnder "äußerlicher" Validität rechtfertigen.
Er fordert also im Dienste der Objektivität eine Priorität linguistischer Valididät über der sog. 'Face Validity'.
Im Gegensatz dazu fordert DAVIES (1977:50), daß auch nicht objektive Tests, wie z.B. Essays, als Item nicht aus einem objektiven Test ausgeschlossen zu sein brauchen. Sie können vielmehr als 'openended questions' betrachtet werden, die objektiv bewertet werden müßten. Dadurch wird das Problem der Objektivität jedoch nur eine Stufe weitergeschoben, nämlich zur Auswertung. Ob dort eine Objektivität erreicht werden kann oder nicht, soll hier nicht diskutiert werden. Der Vorschlag kann aber als Ansatzpunkt gesehen werden, die Grenzen der Überprüfung linguistischer Einzelfertigkeit zu überwinden. Es sei hier vermerkt, daß dieses Dilemma,zwischen 'Face Validity' und Objektivität wählen zu müssen, kennzeichnend ist für eine Testtheorie, die aufbaut auf einer strukturalistischen Linguistik, bei der sprachliche Elemente losgelöst von ihrem Kontext gesehen werden.

Zusammenfassend kann folgendes festgestellt werden:
Objektivität wird gesehen als Vorbedingung für eine stabile quantitative Aussage über das fremdsprachliche Verhalten der Probanden. Je nach Phase der Testdurchführung werden unterschieden:

Durchführungs-, Auswertungs- und Interpretationsobjektivität.

Das Kriterium wird nicht bei allen Autoren explizit genannt, sondern vor allem bei den amerikanischen Autoren in der Wirkungsweise auf die anderer Gütekriterien dargestellt.
Bei Valette, Harris und Heaton wird die Reliabilität beeinflußt, bei Lado zusätzlich die Oberflächenvalidität. Grundsätzlich muß gesehen werden, auch wenn dies nicht immer expliziert dargestellt wird, daß die Objektivität die erste Vorbedingung für die Reliabilität und diese wiederum für die Validität eines Tests ist.

2.1.2.3. Reliabilität

Bevor Testergebnisse untersucht werden können, ist es notwendig zu bestimmen, wie zuverlässig sie sind. Obwohl die Wichtigkeit der Testgültigkeit unbestritten höher einzustufen ist, ist es wesentlich, zuerst ihre Zuverlässigkeit oder Reliabilität zu etablieren.

Lienert versteht unter der Reliabilität eines Testes

> "... den Grad der G e n a u i g k e i t , mit dem er ein bestimmtes Persönlichkeits- oder Verhaltensmerkmal mißt, gleichgültig, ob er dieses Merkmal auch zu messen beansprucht." Lienert (1967:14)

Die Genauigkeit kann gestört werden durch extrinsische und intrinsische Faktoren. Unter ersteren versteht man die Fluktuationen bei der Testdurchführung, Auswertung und/oder Interpretation, die entweder vom Durchführenden oder vom Untersuchten verursacht werden. Sie sind unter dem Begriff der Objektivität dargestellt als Voraussetzungen der Reliabilität (Lado, Valette, Harris und Davies).
Intrinsische Faktoren sind schwerer zu erschließen; es sind die durch den Test selbst bedingten Faktoren der Teststabilität und Testkonsistenz. Ingram (1977:16) unterscheidet deshalb zwischen 'Stability Reliability' und 'Equivalence Reliability'. Ein Testinstrument ist dann stabil, wenn es bei den gleichen Probanden bei zeitlich aufeinanderfolgenden Messungen bei jedem Meßvorgang das gleiche Ergebnis liefert. Dies ist im psychometrischen Bereich sehr unwahrscheinlich[9a] Es genügt aber, wenn die gleiche Rang- und Zuordnung der Probanden untereinander gemessen wird, um eine hohe Stabilitätsreliabilität zu demonstrieren. Zwei oder mehrere Meßinstrumente besitzen dann eine 'Equivalence Reliability', wenn sie, an den gleichen Probanden angewendet, das gleiche Ergebnis liefern.

Die theoretische Grundlage hierfür sieht Ingram wie folgt:

> "The underlying notion is that there is a hypothetical pool of items consisting of all those items which measure the characteristic under investigation, and none that do not, and further that the items which make up an actual test are drawn from that pool. If the items that go to make up Test A have been drawn from the right pool, and the items that go to make up Test Bhave also been drawn from the same pool, then the results from Test A must be equivalent to the results from Test B, and indeed to any other test similarly made up." Ingram (1977:16,17)

Gronlund (1976:108) expliziert neben den genannten noch eine weitere Art von Reliabilität: die 'Internal Consistency' ('innere Konsistenz'); sie gibt die Konsistenz eines Tests über verschiedene Testteile an.

Entsprechend den verschiedenen Aspekten der Reliabilität gibt es verschiedene operationale Zugänge:

1) Die R e t e s t R e l i a b i l i t ä t wird nach der Wiederholung desselben Tests bei derselben Stichprobe durch Berechnung der Korrelation der beiden Testergebnisse ermittelt. Der so gefundene Reliabilitätskoeffizient gibt die Stabilität des Tests über einen Zeitraum an. Lienert gebraucht die Termini 'Merkmalskonstanz' oder - negativ bestimmt - 'Funktionsfluktuation' (^{2}1967:234)

9a Vergleiche Lienert (1967^2:215)

2) Die Paralleltest Reliabilität wird ermittelt, indem die Korrelation von zwei einander streng vergleichbaren Tests (Paralleltests), die der gleichen Stichprobe vorgelegt wurden, berechnet wird. Dieser Reliabilitätskoeffizient gibt an, wie sehr beide Tests tatsächlich dasselbe messen, d.h. wie konstant er bezüglich seiner Bedingungen ist. In der Praxis gibt es viele Fälle, in denen aus verschiedenen Gründen diese Methoden nicht möglich sind.

3) Die Split-half oder Testhalbierungsreliabilität oder Testhalbierungskonsistenz ist jene Methode, die darin besteht, einen einmal durchgeführten Test in zwei gleiche Teile aufzuteilen.

Lienert gibt verschiedene Halbierungstechniken an:

a) die Halbierung nach geradzahligen und ungeradzahligen Aufgaben, die in der Mehrzahl der Fälle gewählt wird;

b) die Halbierung nach Zufall, die systematische Fehler vermeiden helfen soll. Sie wird selten und - wenn überhaupt - nur bei schwierigkeitsgleichen Aufgaben eines Niveautestes angewandt;

c) die Halbierung aufgrund der Analysedaten. Dabei werden systematisch Aufgabenpaare von annähernd gleicher Schwierigkeit und gleicher Trennschärfe miteinander verglichen. Die rechnerische Auswertung geht wie bei a vor sich.

d) Die Halbierung nach Testzeit kommt dann in Frage, wenn der Schnelligkeitscharakter bei einem Test überwiegt.

Die Testergebnisse der beiden Hälften werden korreliert, und der Reliabilitätskoeffizient so aufgewertet, daß er für den ganzen Test Geltung beanspruchen kann.

Die übliche Methode der Aufwertungsberechnung ist die SPEARMAN BROWN - Formel zur Schätzung der Konsistenz des ganzen Testes aufgrund seiner Hälften.

$$r_{tt} = \frac{2 \cdot r_{12}}{1 + r_{12}}$$

Dabei ist r_{tt} die Gesamtkonsistenz

r_{12} die Korrelation von der 1. zur 2. Testhälfte.

Unterschiede können auf verschiedene Aufgabenanforderungen in den beiden Testhälften zurückgeführt werden.

4) Bei der Reliabilitätsbestimmung durch die Konsistenzanalyse werden die Elemente eines Tests als mehrfach halbierte Testteile aufgefaßt. Über die Nennwerte der Aufgabenschwierigkeit und der

Trennschärfe sämtlicher Aufgaben wird auf indirektem Wege der Reliabilitätskoeffizient bestimmt. Das statistsiche Hilfsmittel hierzu ist die bekannte KUDER-RICHARDSON-20-FORMEL (siehe Lienert,[2] 1967:227).

$$r_{tt} = \frac{n}{n-1}\left[\frac{s_x^2 - n\bar{p}\bar{q}}{s_x^2}\right] = \frac{n}{n-1}\left[\frac{s_x^2 - \Sigma pq}{s_x^2}\right]$$

Dabei ist s_x^2 = Varianz der Test-Rohwerte X

p = Schwierigkeitsindex als Verhältniszahl ausgedrückt P/100

q = (1 - p)

Die Testhalbierungsmethode und die Konsistenzanalyse (3 und 4) unterscheiden sich wesentlich von der Retest- und Paralleltestmethode. Sie wurde in vielen Fällen bei der Erforschung des CT verwendet (siehe Tafel 16a). 3 und 4 werden nach Lienert, der den terminologischen Empfehlungen der APA (1954) folgt, der 'inneren Konsistenz' eines Testes zugeordnet (siehe oben Gronlund). Als präzise Schätzformeln gibt Lienert ([2]1967:223) folgende Formel von KRISTOF an.

$$r_{tt} = \frac{2}{N-1} + \frac{N-3}{N-1} \cdot \frac{4s_1 s_2 r_{12}}{s_1^2 + s_2^2 + 2s_1 s_2 r_{12}}$$

Die Form der Konsistenzanalyse ist die angemessenste Art der Reliabilitätsbestimmung, wenn ein Test nur eine Form hat und nur einmal angewendet werden kann. Insbesondere bei Sprachtests, bei denen Gedächtnisfaktoren eine entscheidende Rolle spielen und bei denen keine vollständige Aufgabenanalyse durchgeführt werden kann, bietet sie sich als geeignetes Mittel an (siehe S. 116).

S t ö r f a k t o r e n der R e l i a b i l i t ä t

Im folgenden Abschnitt werden eine Reihe von Störfaktoren aufgezählt, die die Reliabilität beeinflussen.[10]

1) Der H o m o g e n i t ä t s g r a d eines Tests

Homogenität sei als inhaltliche Einheitlichkeit beschrieben, d.h. ein homogener Test mißt nur ein eng umschriebenes Persönlichkeitsmerkmal. Homo-

10 Eine ausführliche Darstellung findet sich bei Lienert (1967:236)

gene und heterogene Tests müssen hinsichtlich der verschiedenen Reliabilitätsfaktoren unterschiedlich eingestuft werden. Dabei besitzen homogene Tests hohe innere Konsistenz, heterogene Tests lassen in der Regel keinen hohen Konsistenzkoeffizienten erwarten. Den höchsten und zugleich sinnvollsten Reliabilitätskoeffizienten bei heterogenen Tests bringt die Testwiederholung. Eine Reliabilitätsverbesserung heterogener Tests kann auch durch eine Aussonderung heterogener Aufgaben erreicht werden. Lienert sieht die Problematik heterogener Tests und stellt fest:

> "Die geringen Möglichkeiten und die Unsicherheit der Reliabilitätskontrolle bei heterogenen Testen könnten manchen zu dem Schluß verleiten, daß man nur homogene Tests aufbauen sollte. Das trifft jedoch keineswegs zu, da heterogene Teste ... mit praktischen Validitätskriterien oft höher korrelieren als homogene Teste." (1967:236)

Es ist deshalb darauf zu achten, inwiefern Sprachtests heterogen oder homogen sind. Sind sie so angelegt, daß verschiedene Skills getrennt geprüft werden, können homogenere Einzeltests konstruiert werden, als wenn e i n Test verschiedene Fertigkeiten auf einmal prüft. Wird nun gefordert, die Reliabilität - die also bei heterogenen Tests niedriger ist - durch Testwiederholung zu erhöhen, bringt dies ein schwer lösbares Problem mit sich, denn Sprachtests sind wegen der großen Bedeutung des Gedächtnisfaktors kaum wiederholbar. Man wird bei solchen Tests also mit Reliabilitätsstörungen rechnen müssen und/oder andere Möglichkeiten zur Reliabilitätsverbesserung, wie z.B. eine Testverlängerung, anwenden müssen.

2) Die S c h n e l l i g k e i t s k o m p o n e n t e ist bei Schnelligkeitstests zu berücksichtigen.

Wird ein Sprachtest als 'Wissens- und Leistungstest' - etwa um die "Menge" gelernten "Materials" zu erfassen - gesehen, hat diese Komponente sicher einen Einfluß auf die Testreliabilität. Wird aber als Lern- und Testziel die Kommunikationsfähigkeit gesehen und die Schnelligkeit als irrelevant in normalen Kommunikationssituationen erkannt, kann diese Komponente vernachlässigt werden.

3) Die i r r e p r ä s e n t a t i v e V a r i a b i l i t ä t ist ein weiterer Störfaktor, von dem Lienert dann spricht, wenn die für die Kontrollstichprobe benutzte Probandengruppe eine geringere oder größere Streuung hat als die Population, für die der Test gedacht ist. Das Korrelationsdiagramm (Abb. 1) zeige etwa die Reliabilität einer Testform A in Korrelation zu Testform B.

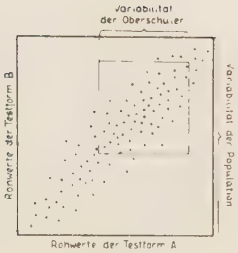


ABB 1

(entnommen aus Lienert,[2] 1967:238)

Der Ausschnitt für die Variabilität der Stichprobe der Oberschüler läßt deutlich erkennen, daß deren Reliabilität im Vergleich zur Gesamtpopulation irrepräsentativ ist. Lienert erklärt:

"Da man für den Zweck der Reliabilitätskontrolle zuerst Stichproben heranzieht, die sich einem bequem anbieten und die in der Regel soziologisch homogener sind, besteht oft die Vermutung zu Recht, daß bei diesen auch das untersuchte Persönlichkeitsmerkmal nicht über der gesamten Populationsbereich streut." (1967:238) Diese Vermutung und Unstimmigkeit wird oft erst nach der Eichung offenbar, da diese erst die Populationsvarianz erkennen läßt.

Wie das Beispiel zeigt, ist vor allem im schulischen Bereich - wo häufig kleine Probandengruppen akzeptiert werden müssen,- mit diesem Störfaktor zu rechnen. (Siehe TAFEL 31)

4) Die T e s t l ä n g e kann als vierter Störfaktor bezeichnet werden. Im allgemeinen sind längere Tests unter sonst vergleichbaren Bedingungen reliabler als kürzere. Testlänge darf hier nicht verwechselt werden mit der zur Lösung benötigten Zeit, sondern bezieht sich auf die Anzahl der Aufgaben. Der typische Test besteht aus einer Anzahl separat gewerteter Aufgaben, aus deren linearer Kombination sich das Testergebnis ergibt. Soll die Testreliabilität nun verbessert werden, so können Aufgaben hinzugefügt werden.

Ähnlich wie bei 1) ergibt sich auch hier das Problem, daß Sprachtests, die aus einer Reihe von separaten Einzelaufgaben bestehen, eher dem 'typischen Test' entsprechen und somit in ihrer Reliabilität weniger störanfällig sind als solche, die mehrere Fähigkeiten integrieren. Aus Einzelaufgaben bestehende Tests dürften in der Regel auch relativ leicht um Zusatzaufgaben zur Verbesserung ihrer Reliabilität zu erweitern sein.

5) Der S t r e u u n g s b e r e i c h der A u f g a b e n - s c h w i e r i g k e i t ist ein weiterer Störfaktor. Je mehr Aufgaben von mittlerer Schwierigkeit ein Test enthält, umso höher ist unter sonst gleichen Bedingungen seine Reliabilität. Diese Tatsache sieht von einem anderen Standpunkt so aus: je weiter eine Aufgabe vom mittleren Schwierigkeitsgrad entfernt ist, umso weniger funktioniert sie differenziell. Ist eine Aufgabe zu leicht, werden sie alle Probanden lösen, ist sie zu schwer, kann sie keiner lösen.
Hier gilt für Sprachtests wiederum Ähnliches wie bei 1) und 4), denn aus Tests, die aus Einzelaufgaben bestehen, sind leicht Aufgaben unerwünschter Schwierigkeit zu entfernen, während dies bei den sog. 'integrativen Tests' (siehe 2.1.3.4.) nicht ohne eine unter Umständen grundlegende Testveränderung möglich ist.

6) Der Z u f a l l s f a k t o r bei der Aufgabenbeantwortung vermindert die Chancen für eine hohe Reliabilität ein wenig.
Enthält ein Sprachtest offene Antwortmöglichkeit ist dieser Störfaktor in Betracht zu ziehen. Dies ist in der Regel seltener der Fall bei Tests, die aus Einzelaufgaben bestehen, und häufiger bei sog. 'subjektiven Tests'. Ein Testverfahren wie das des CT reduziert trotz gewisser offener Antwortmöglichkeiten den Zufallsfaktor (siehe 2.4.).

7) Einzelne allzu hohe Aufgaben-Interkorrelationen können die Reliabilität vermindern, indem sie durch gemeinsame richtige oder falsche Antworten den Test gleichsam verkürzen (siehe auch Punkt 1).
Auch hier setzt sich die bei 1), 4) und 5) aufgezeigte Problematik fort. Sprachtests aus Einzelaufgaben können die Interkorrelation von Aufgaben leichter vermeiden als 'integrative Tests'.

Insgesamt kann gesagt werden, daß die Reliabilität von aus Einzelaufgaben bestehenden Tests weniger störanfällig ist als die von sog. 'integrativen Tests'.

D i e A u f g a b e n a n a l y s e

Die Reliabilität eines Tests wird nach Lienert (^{2}1967:70) auch noch von den Gütekriterien der einzelnen Aufgaben beeinflußt. Diese sind: Schwierigkeit, Trennschärfe und Validität.
Der S c h w i e r i g k e i t s i n d e x einer Aufgabe ist definiert durch die prozentuale Häufigkeit, mit der sie von einer repräsentativen Stichprobe von Probanden beantwortet wird. Nach dieser Definiton muß die Formel für den Schwierigkeitsindex (nicht bei Schnelligkeitstests!) lauten:

$$P = 100 \cdot \frac{N_R}{N}$$

Dabei ist N_R die Anzahl der Probanden, die die Aufgabe richtig beantwortet haben

N die Gesamtzahl der Probanden

Brauchbare Aufgaben befinden sich im Bereich zwischen 0,3 und 0,7, wobei 0,3 als schwer und 0,7 als leicht zu bezeichnen sind.
Die Testreliabilität kann erhöht werden, indem Aufgaben, die zu schwer oder zu leicht sind, ausgeschieden werden. (Vergleiche Punkt 5 der Störfaktoren und der Reliabilität).

Der Trennschärfenkoeffizient einer Aufgabe ist definiert durch den biserialen[10 a] Korrelationskoeffizienten zwischen der Aufgabenbeantwortung der Probanden (ja/nein; richtig/falsch) einerseits und den von ihnen erzielten Rohwerten im Gesamttest andererseits.
Auch Aufgaben, die nicht trennscharf genug sind, sollen ausgeschieden werden, um die Reliabilität zu erhöhen.

Der Validitätskoeffizient wird seltener verwendet. Er ist definiert durch den biserialen oder tetrachorischen Korrelationskoeffizienten zwischen der Aufgabenbeantwortung einerseits und dem Grad der Merkmalsausprägung - gemessen an einem Außenkriterium - andererseits.

Außer den genannten, teilweise quantifizierbaren Kriterien, ist für die einzelne Aufgabe auch ihre inhaltliche Gestaltung von Bedeutung, d.h.u.a.

> "Jede Aufgabe soll von jeder anderen inhaltlich unabhängig sein. Die Lösung einer Aufgabe darf keinen Hinweis für die Lösung irgend einer anderen enthalten und schon gar nicht deren Lösung bedingen." (Lienert[2] 1967:41)

Die beiden Störungsfaktoren der Reliabilität von Sprachtests dargestellte Problematik erscheint auch hier wieder bei der Aufgabenanalyse.
Während bei Sprachtests, die aus Einzelaufgaben bestehen, die Anaylse dieser Aufgaben leicht möglich ist und die eventuelle Änderung einzelner Aufgaben zur Erhöhung der Testreliabilität als Ziel hat, ist bei integrativen Tests die Einzelaufgabe kaum veränderbar, ohne den Charakter des Tests zu verändern, weshalb ihre Durchführung in Frage gestellt werden muß.

---

10 a Die Ausdrücke biserial und tetrachorisch sind nur gebraucht zur Darstellung der statistischen Verfahrensweise; sie sind aber für das Verständnis der Testauswertung nicht unbedingt notwendig. Eine Gradangabe der Temperatur erfordert ja vom Nichtphysiker auch kein absolutes Verstehen der physikalischen Definition!

Interpretation von Reliabilitätswerten

Die verschiedenen Arten der Reliabilitätsbestimmung ergeben numerisch unterschiedliche Werte. Im Hinblick auf Unterschiede von Testinhalt und Verwendungszweck ist es schwierig, konkrete Größenangaben für den Reliabilitätskoeffizienten zu machen.
Ingenkamp hält einen Wert von 0,7 für zufriedenstellend, wenn es gilt, den Leistungsstand einer Gruppe zu ermitteln (1973[3]:65). Bei Lienert genügt hier schon ein Wert von $r_{tt} = 0,5$ (1967[2]:21). Für die Beurteilung individueller Differenzen ist ein Wert von $r_{tt} = 0,7$ eben noch ausreichend. Standardisierte Tests sollten eine innere Konsistenz von $r_{tt} = \geq 0,9$ und eine Retest- oder Paralleltestreliabilität von $\geq 0,8$ aufweisen (Ingenkamp, 1967:309). Keine Interpretation sollte erfolgen ohne Einbeziehung aller die Reliabilität beeinflussenden Störfaktoren, des Testzweckes sowie der Angaben der Berechnungsmethoden und statistischen Werte, um den größtmöglichsten Nutzen daraus ziehen zu können.

Beim Vergleich von Reliabilitätskoeffizienten muß der Interpret das Gewicht der verschiedenen Faktoren gegeneinander abwägen, in die Interpretation einbeziehen und mit ähnlichen auf dem Gebiet vorliegenden Ergebnissen vergleichen. Das Ergebnis ist dann zufriedenstellend, wenn die sonst üblichen Werte überschritten werden, es ist unbefriedigend, wenn sie unterschritten werden.

Résumée: Die Reliabilität eines Tests kann unter mehreren Aspekten gesehen werden. Vier Arten wurden erkannt, die jeweils verschiedene methodische Zugänge erfordern: Retestreliabilität, Paralleltestreliabilität, Split-Half-Reliabilität (Testhalbierungsreliabilität) und Konsistenzanalyse. Extrinsische und intrinsische Faktoren beeinflussen den Reliabilitätsgrad: Objektivität, Homogenität, Testlänge, Repräsentativität der Variabilität der Stichprobe, Streuungsbereich der Aufgabenschwierigkeit, Zufallsfaktoren bei der Aufgabenbeantwortung und Aufgabeninterkorrelation. Zur Interpretation der Reliabilitätskoeffizienten sind alle relevanten Faktorenwerte einzubeziehen.
Im Bereich von Sprachtests sind die Testhalbierungsreliabilität und die Konsistenzanalyse vorzuziehen. Die Reliabilität von integrativen Tests, wie z.B. dem CT, ist eher störanfällig als die von 'discrete-point-tests', bei denen Störfaktoren leicht ausgeschaltet oder verringert werden können.

2.1.2.4. Validität

Validität wird übereinstimmend im weitesten Sinn definiert als der Grad der Genauigkeit, mit der ein Test tatsächlich das Verhaltensmerkmal mißt, das er vorgibt zu messen, bzw. das er messen soll.

Lado (1961:321) bezeichnet Validität als die Relevanz zwischen dem Meßvorgang und dem Meßziel. Mehrere Möglichkeiten, diese Relevanz zu messen bzw. zu bestimmen, führen zur Unterscheidung verschiedener Aspekte der Validität, wobei die Terminologie nicht einheitlich ist.

Lienert (^{2}1967:16 ff und 255 ff) unterscheidet zwischen

1) inhaltlicher Validität (content validity)
2) Konstruktvalidität (construct validity)
3) kriteriumsbezogener V. (bei Gronlund ist dafür der englische Terminus 'criterion-related validity' zu finden) Gronlund, 1968:107-109

Für 'inhaltliche Validität' wird auch der Begriff 'logische Validität' ('logical validity') verwendet; z.B. bei Dieterich (1972:75).

Die 'kriteriumsbezogene Validität' wird von Ingram als 'pragmatic validity' bezeichnet. Lado, Dieterich, Heaton u.a. verwenden dafür den Begriff 'empirical validity' ('empirische Validität').

Beide Begriffe scheinen nicht sehr zweckmäßig zu sein, da auch die anderen Arten der Validität auf einen pragmatischen, bzw. empirischen Basisbereich zu reduzieren sind.

Gelegentlich wird der Begriff 'kriteriumsbezogene Validität' überhaupt nicht aufgeführt, sondern die zwei Unterarten, nämlich 'Übereinstimmungsvalidität' ('concurrent validity') und 'Vorhersagevalidität' ('predictive validity') eigenständig verwendet. Die Unzulänglichkeit des Begriffs der 'kriterienbezogenen Validität' beruht darin, daß auch alle anderen Validitätsarten in Beziehung zu einem Kriterium dargestellt werden können, wie dies bei Davies (^{2}1970:9) geschieht:

TAFEL 3

T y p e	related to	C r i t e r i o n
Face	⟶	Layview of language
Content	⟶	Syllabus analysis
Construct	⟶	Theory of aims in language teaching
Concurrent	⟶	Another test or measure (e.g. teacher' s assessment) of language ability given at some time

Einige Autoren (Davies, 1970 und Davies/Allen, 1977; Heaton, 1975; Ingram, 1978; Lado, 1961) führen noch die sog. 'face validity' als weitere, besonders im Bereich der Fremdsprachentests bedeutsame Validitätsart auf. Der Begriff wird von Freudenstein (1971) mit 'Gültigkeit durch Augenschein' übersetzt. Lienert (^{2}1967) verwendet dafür den Begriff 'psychologische Gültigkeit'.

I n h a l t l i c h e V a l i d i t ä t

Wenn sich der Testinhalt und das zu messende Persönlichkeitsmerkmal decken, ist ein Test inhaltlich valide. Diese Möglichkeit, einen Test sozusagen per definitionem zu validieren (z.B. HET 6 : "Dieser Test besitzt Inhaltsvalidität. Die Testaufgaben repräsentieren das Ziel des Testes." Zitiert nach Fettweiß et al (1966:8)), sollte - wie Lienert ([2]1967:260) fordert - durch das 'Rating' einer Gruppe kompetenter Beurteiler objektiviert werden.

Verhältnismäßig leicht ist die inhaltliche Validität bei Kenntnistests oder bei engumgrenzten Lernzielkontrollen festzustellen. Im Falle von Tests einer allgemeinen Fremdsprachenkompetenz ist diese Aufgabe nicht so eindeutig zu lösen: Es ist festzustellen, wie repräsentativ eine Aufgabenauswahl für die Fremdsprachenkompetenz ist. Hierzu ist eine axiomatische Vorentscheidung nötig, nämlich was unter 'Fremdsprachenkompetenz' zu verstehen ist (siehe 2.2.).[10b]

Die inhaltliche Validität eines Testes ist numerisch nicht quantifizierbar, sondern kann bestenfalls aus dem 'Rating' der Beurteiler erkannt werden.

Es bietet sich an dieser Stelle an, den Begriff der F a c e V a l i d i t y vorzustellen, denn er steht in engem Zusammenhang zur inhaltlichen Validität. Auch die Face Validity bezieht sich auf den Inhalt eines Testes. Während aber die inhaltliche Gültigkeit eine genaue linguistische und pragmatische Untersuchung fordert, bezieht sich die Face Validity nur auf die Glaubwürdigkeit und Akzeptabilität eines Tests, primär vom Standpunkt des Probanden aus.
Ingram bezeichnet es als " a public relations problem rather than a technical one" (1978:18). Wiederum ist eine Abhängigkeit zur grundsätzlichen Sichtweise des Phänomens Sprache erkennbar: Wird Sprache nur als habitualisiertes Verhalten gesehen,ist der Aspekt zu vernachlässigen. Wird aber Sprache als grundsätzlich kontextabhängiges, soziales Phänomen betrachtet, erhöht sich der Stellenwert der Face Validity. Es ist zu erwarten, daß ein Proband durch einen glaubwürdigen, psychologisch gültigen Test anders motiviert wird und anders auf ihn reagiert, als auf einen, dessen Aufgaben und Ziel nicht eindeutig einsichtig sind.

10b Eine ausführliche Darstellung der Repräsentativität von Testaufgaben für die Fremdsprachenkompetenz - sowohl bei integrativen als auch bei 'discrete-point-tests' findet sich in 2.1.3.

Während die inhaltliche Validität auf den sprachlichen Inhalt eines Testes bezogen ist, ist die psychologische Analyse des Testes das Ziel der

K o n s t r u k t v a l i d i t ä t .

Unter 'Konstrukt' wird hier in Anlehnung an Dieterich (1972:76) ein wissenschaftlicher Terminus verstanden, der nicht direkt operationalisierbar ist, sondern durch Einbettung in ein Netz von Einzelaussagen, Einzelereignissen, quantitativen und qualitativen Informationen und theoretischen Folgerungen abgegrenzt wird.
Das Verfahren der Konstruktvalidierung besteht in der theoretischen Klärung dessen, was der betreffende Test mißt. Hierzu werden allerdings neben einer logischen Analyse auch empirisch statistische und experimentelle Methoden herangezogen.

Lienert (^{2}1967:262) gibt eine ganze Reihe solcher experimenteller Methoden an: Ihre Anwendung hängt von der Sichtweise des Konstrukts der Fremdsprachenkompetenz ab (cf. 2.2.):

"1) Korrelation des Testes mit Außenkriterien;
2) Korrelation des Testes mit Testen ähnlichen Validitätsanspruches;
3) Korrelation mit Testen, die andere Persönlichkeitsmerkmale erfassen;
4) Faktorenanalyse des zu validierenden Testes gemeinsam mit Außenkriterien, validitätsverwandten und validitätsdivergenten Testen;
5) Analyse Individueller Unterschiede in den Testresultaten;
6) Analyse intraindividueller Veränderungen bei wiederholter Durchführung mit und ohne systematische Variation der Durchführungsbedingungen;
7) inhaltlich-logische Analyse der Testelemente."

Im Gegensatz zur inhaltlichen Validierung, die numerisch überhaupt nicht erfaßbar ist, können hier empirisch gefundene numerische Werte in die Begründung einfließen und so die Objektivität steigern.

K r i t e r i u m s b e z o g e n e V a l i d i t ä t besitzt ein Test dann, wenn ein empirisch nachweisbarer Zusammenhang besteht zwischen Test und einem Kriterium. Das Kriterium muß in einer direkten oder indirekten Weise das zu erfassende Persönlichkeitsmerkmal erfassen oder widerspiegeln. Das Kriterium erhält per definitionem selbst Validität.

Lienert (^{2}1967:25) unterscheidet Gruppierungen von Kriterien, die verschiedene Arten kriterienbezogener Validität liefern:

a) innere - äußere kriterienbezogene Validität
b) Übereinstimmungs- und Vorhersagevalidität

ad a) Eine objektiv bewertete Kriteriumsleistung oder ein Schätzurteil, konkret etwa eine Lehrerbeurteilung der Fremdsprachenkompetenz, wird als äußeres Kriterium bezeichnet und zur Bestimmung der äußeren Validität

verwendet. Ein bereits für dasselbe Kriterium validierter Test wird als inneres Kriterium bezeichnet und zur Bestimmung der inneren Validität verwendet. Es ist einleuchtend, daß ein solcher Validitätskoeffizient "aus zweiter Hand" niedriger liegen wird, als ein solcher aus "erster Hand". Um Fehlinterpretationen zu vermeiden, wird empfohlen, sich mehr auf den Grad der logischen Validität zu stützen und/oder andere Kriterien heranzuziehen.

ad b) Sind Kriterienwert und Testergebnis zeitlich koexistent vorhanden, spricht man von Übereinstimmungsvalidität ('concurrent validity'). Ist das Kriterium erst in der Zukunft verfügbar, spricht man von Vorhersagevalidität ('predictive validity'). Prognostische Tests bedürfen einer Vorhersagevalidierung. Für die vorliegende Untersuchung dienten Schulnoten und Lehrerurteil als Kriterien.

Die kriteriumsbezogene Validität wird von Lienert historisch und praktisch gesehen als der bedeutsamste Aspekt der Validität bezeichnet. Gegenüber den anderen Validitätsarten hat sie den Vorzug, quantitativ numerisch erfaßbar zu sein.
Die Methode ihrer Bestimmung ist in der Regel das als PEARSON-PRODUKT-MOMENT-KORRELATION bekannte Verfahren. Die Höhe der nötigen Validitätskoeffizienten hängt von seinem speziellen Verwendungszweck ab. Lienert gibt folgende Orientierungshilfe:

> "Soll ein bestimmter Proband möglichst treffend hinsichtlich des Ausprägungsgrades eines Persönlichkeitsmerkmales oder seiner Eignung beurteilt werden, so sind hierfür Tests mit einer $r_{tc} > 0{,}7$ erforderlich, vor allem, wenn das Urteil über den Probanden allein oder fast allein vom Testergebnis abhängt.
> In der Testpraxis werden so hohe Validitätskoeffizienten nur ausnahmsweise bei Einzeltesten erzielt. Testbatterien - also Kombinationen von Einzeltesten - haben in dieser Hinsicht weit bessere Chancen, die Anforderungen zu erfüllen.
> Wenn - wie dies in der diagnostischen Praxis die Regel ist - neben dem Testergebnis noch andere Informationen für die individuelle Eigenschafts- oder Eignungsbeurteilung herangezogen werden, so reduziert sich die Anforderung an den Validitätskoeffizienten je nach der Gewichtigkeit dieser Informationen auf Werte von $r_{tc} = 0{,}5$.
> Bei der individuellen Begutachtung gilt das Interesse des Psychologen ausschließlich dem jeweiligen Pb. Anders liegt der Fall bei der Eignungsauslese." Lienert (²1967:310)

Allgemein gesprochen soll ein Test so valide sein, daß seine Anwendung eine bessere Voraussage möglich macht als seine Unterlassung. Man wird sich prin-

zipiell mit jedem Test zufrieden geben müssen, der sich statistisch auf der 1%-Stufe sichern läßt, der also einen wahren Zusammenhang zwischen Test und Kriterium zeigt. Oft sind Tests mit Validitätskoeffizienten von 0,3 von praktischer Bedeutung.
Auch die Art des gemessenen Persönlichkeitsmerkmals, ob es z.B. durch Testverfahren schwer oder leicht zu erfassen ist, beeinflußt die Beurteilung der Validitätskoeffizienten.

Zusammenfassung:

Auch das Gütekriterium der Validität ist unter verschiedenen Aspekten zu sehen. Als drei Hauptarten kristallisieren sich

1) die inhaltliche Validität,

2) die Konstruktvalidität und

3) die kriteriumsbezogene Validität heraus.

Mit ihrer Hilfe können Aussagen über die Relevanz zwischen dem Test und dem zu messenden Persönlichkeitsmerkmal gemacht werden. Direkt numerisch quantifizierbar sind nur die Angaben über die kriterienbezogene Validität, die deshalb von besonderer Wichtigkeit sind. Im Bereich der Fremdsprachenkompetenzmessung ist die Anwendung der einzelnen Validitätsarten durch die Sichtweise der Fremdsprachenleistung mitbestimmt. (Siehe 2.2.)

2.1.2.5. Normierung

Sind für einen Test die Gütekriterien der Reliabilität und Validität zufriedenstellend, kann er im Zuge der Teststandardisierung normiert bzw. geeicht werden. Das heißt, es wird ein Bezugssystem geschaffen, das für die Einordnung verschiedener individueller Testergebnisse dienen kann. Eine Normierung kann für die Gesamtpopulation oder für bestimmte soziale Gruppen erfolgen. (Vgl. Kamratowski et al, 1973:78).

Beziehen wir uns auf die in Matrix 2 dargestellten Testformen des standardisierten Tests, des informellen Tests und der Aufgabensammlung, so stellen diese Formen unterschiedlich hohe Anforderungen an die Gütekriterien:

1) Standardiserte Tests müssen höchste Ansprüche an Validität und Reliabilität erfüllen, damit sie auch überregional etwa zu Zwecken der Eignungsfeststellung eingesetzt werden können.

2) Informelle Tests sind in der Regel vom Klassenlehrer für eine genau eingegrenzte Zielgruppe einer Klasse, gegebenenfalls einiger Parallelklas-

sen, bestimmt.
Für sie genügt eine einfache Normierung im Hinblick auf die Zielgruppe. Das Niveau der übrigen Gütekriterien wird vom Testentwerfer im Hinblick auf seine Zwecke festgelegt, sollte aber so hoch wie möglich liegen.

3) Einfache Aufgabensammlungen, die nur zur Überprüfung spezieller Lernziele dienen sollen, stellen die geringsten Anforderungen an Gütekriterien. Der Durchführende, meist der Lehrer, wird im Hinblick auf die Gewichtung der Arbeit selbst die Höhe der Anforderungen bestimmen.

2.1.2.6. Ökonomie und Testadministration

Soll ein Test ökonomisch sein, muß er folgende Bedingungen erfüllen:

1) kurze Durchführungszeit
2) geringer Materialverbrauch
3) einfache Durchführbarkeit
4) Durchführbarkeit als Gruppentest
5) schnelle und leichte Auswertbarkeit

Die Punkte dieser Liste sprechen für sich selbst. Sie schließen im wesentlichen auch die bis jetzt noch nicht behandelten Kriterien der eingangs genannten Autoren ein. Ganz allgemein ist dazu zu bemerken, daß ein Test, der diese Bedingungen nicht erfüllt und deshalb nicht durchgeführt werden kann, trotz nachgewiesener Objektivität, Reliabilität und Validität wertlos ist.

Wegen des großen Einflusses der Testadministration auf die anderen Gütekriterien soll noch auf einige technische Aspekte der Testdurchführung besonders hingewiesen werden. Der diagnostische Wert eines Testes hängt wesentlich von der Einhaltung objektiver Bedingungen ab. Es ist nötig, diese im voraus festzulegen und sie den Durchführenden und den Probanden mitzuteilen. Folgende Punkte sind zu beachten:

Gestaltung des Testmaterials

Es versteht sich von selbst, daß Lesbarkeit und Übersicht gewährleistet sein müssen. Der Proband muß klar erkennen, wo er seine Antworten einzutragen hat. Dabei ist Vorsorge für möglichst rationelle Auswertungsmöglichkeiten zu tragen (Übersicht, siehe Lienert, ^{2}1967:200 ff).

Testzeit

Abgesehen von 'speedtests' muß dafür gesorgt werden, daß genügend Zeit

zur Lösung zur Verügung steht. Die Zeit soll so bemessen sein, daß alle Probanden alle Aufgaben in Angriff nehmen können, daß aber zur Vermeidung disziplinarischer Schwierigkeiten besonders für die schnellen Schüler nicht zu viel Zeit übrig bleibt. Insgesamt sind kürzere Tests vorzuziehen, um große Unterschied von vorneherein zu vermeiden. Gegebenenfalls sind Vortests nötig, um eine optimale Länge zu erproben.

Testinstruktion

Der Proband hat ein Recht darauf, zu erfahren, wozu getestet wird. Lienert gibt einen Vorschlag, was die Testinstruktion beinhalten soll:

"1. Was ist der Sinn und Zweck des vorliegenden Testes?
2. Welche Konsequenzen hat er für die Teilnehmer?
3. Besteht Teilnahmepflicht oder ist die Teilnahme freiwillig?
4. Wie steht es mit "Raten" und "Abschreiben"?
5. Was wird von dem Pb erwartet und - ebenso - was wird nicht erwartet?
6. Kommt es mehr auf Schnelligkeit oder mehr auf Gewissenhaftigkeit an?
7. Was soll an Schreibutensilien bereit gehalten werden?"
(G. A. Lienert 1967:193)

Diese Vorschläge sind leicht auf die jeweilige Fremdsprachentestsituation zu übertragen. Als Beispiel sei auf 4.7. verwiesen.

Motivation und Anspannung

Im Fremdsprachenbereich besitzen diese Faktoren wesentliche Bedeutung. Es ist wünschenswert, die Motivation des Schülers möglichst hoch zu halten, die Anspannung aber nicht leistungshemmend werden zu lassen und zwar konstant bei jeder Testdurchführung. Es ist sehr schwer, wenn nicht unmöglich, diese Bedingung zu erfüllen, da große interindividuelle Unterschiede in der Reaktion der Pbn auf Motivierungsimpulse bestehen. Wenn irgendwie möglich, sollen aber dahingehende Bemühungen unternommen werden und im Testmanual erwähnt werden.

Äußere Bedingungen

Die Kontrollierbarkeit der äußeren Bedingungen ist unterschiedlich. Relativ leicht konstant zu halten sind die Arbeitsbedingungen wie Sitz- und Schreibgelegenheit, Arbeitsraumgröße, Helligkeit und Temperatur. Schwer zu kontrollieren sind endogene Faktoren der Pbn wie Ermüdung, Wohlbefinden, Mißbehagen etc. Sie erscheinen kaum in der Diskussion, müssen aber als mögliche Fehlerquellen beachtet werden.

2.1.2.7. Die Wechselbeziehungen zwischen den Gütekriterien

Mehrfach wurde bereits auf die Beziehung von einem zum anderen Kriterium hingewiesen. Dies sei nochmals im Überblick zwischen allen Kriterien aufgezeigt:
Objektivität ist die Vorbedingung jeder Reliabilität, die Reliabilität wiederum Vorbedingung für die Validität. Oft wird Reliabilität als Form

der Validität gesehen (Allen/Davies, 1977:58).

Diese Abhängigkeiten sind auch umkehrbar:

> "Ein Test mit einer hohen Validität muß notwendigerweise auch hohe Objektivität, Konsistenz und Zulänglichkeit besitzen. Die Feststellung einer hohen Validität entbindet somit in gewissem Maße von der Überprüfung der übrigen Gütekriterien." Lienert (1967:20)

Lienert hat hier die Retestreliabilität ausgenommen. Man kann sich vorstellen, daß ein Persönlichkeitsmerkmal sehr treffsicher erfaßt wird, daß es aber zeitlich sehr fluktuiert.

Auf die überragende Bedeutung der Validität gegenüber dem Testinhalt sei noch einmal hingewiesen mit folgendem Zitat von Ingram (1977:12):

> "Once the validity has been demonstrated, however, it is immaterial what type of test we are dealing with. If a test works for the purpose it was intended to, then that is all that matters."

Wegen seiner Wichtigkeit bei der Bestimmung der Reliabilität und Validität soll der Begriff der 'Korrelation' hier kurz als Orientierungshilfe ohne statistisch mathematische Ausführlichkeit dargestellt werden:

Unter 'Korrelation' versteht man die Tendenz des Kovariierens der Testergebnisse eines Testes einer Gruppe von Probanden mit den Testergebnissen eines zweiten Testes der gleichen Gruppe, d.h. daß die beiden Ergebnisse in ähnlicher Richtung und Größe variieren.

Wenn z.B. die Probanden, die eine hohe Punktzahl in einem CT erzielen ebenso hohe Werte in einem anderen erreichen, und Probanden, die niedrige Werte in einem CT erzielen ebenso niedrige Werte in einem anderen Test erreichen, korrelieren die beiden Tests miteinander.

Wenn die Ergebnisse beider Tests genau proportional zueinander verlaufen, ist die Korrelation perfekt, der Korrelationskoeffizient beträgt dann $r = 1{,}00$; besteht kein Zusammenhang beträgt der Korrelationskoeffizient $r = 0$.

Eine Korrelation kann auch negativ sein. Wenn Probanden, die hohe Testergebnisse erzielen, mit niedrigen Punktezahlen bewertet werden und bei niedrigen Testergebnissen mit hohen Punktzahlen - wie das durch unser Zensurensystem geschieht -, sind Korrelationen negativ. Perfekte Proportionalität der Werte ergibt eine perfekte negative Korrelation von $r = -1{,}00$.

Graphisch dargestellt sieht das etwa so aus:

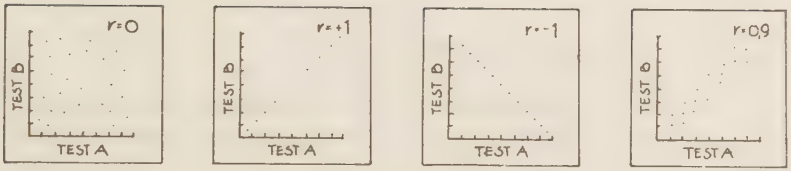


ABB 11

Ein Korrelationskoeffizient von r = 0,9 bedeutet keineswegs eine 90%ige Übereinstimmung.. Eine solche Darstellung verbietet sich für r, kann jedoch für r^2 vorgenommen werden, d.h. daß bei einem Korrelationskoeffizienten von r = 0,9 eine gegenseitige Übereinstimmung der Testergebnisse von r^2 = 0, 81%.

r^2 wird auch als gemeinsame Varianz oder Determinationskoeffizient bezeichnet und gibt an, in welchem Ausmaß die Varianz der anderen Variablen bestimmt wird, oder daß beim obigen Beispiel 81% der Information des einen Tests auch vom anderen Test abgedeckt ist. Während hohe Korrelationen leicht interpretiert werden können, sind niedrige Korrelationen weniger informativ. Niedrige Korrelationen können mehrere Ursachen haben: beide oder einer der Tests sind zu schwer oder zu leicht; einer oder beide Tests sind nicht reliabel; einer oder beide Tests messen nicht, was sie messen sollen, bzw. sind nicht valide.

Zusammenfassung:

Die Qualität eines Tests ist durch mehrere Gütekriterien bestimmt, die sich gegenseitig beeinflussen bzw. bedingen: Objektivität, Reliabilität, Validität und Ökonomie.

Ökonomie und Objektivität sind von außen direkt erfaßbar und in bestimmtem Umfang kontrollierbar.

Die Reliabilität kann nur durch statistische Methoden erfaßt werden. Dabei gibt es mehrere Wege, die verschiedenen Arten der Reliabilität zu erschließen, welche unterschiedliche Aussagen machen.

Auch die Validität kann unter verschiedenen Aspekten gesehen werden. Hauptarten sind die inhaltliche Validität, die Konstruktvalidität und die kriteriumsbezogene Validität. Letztere ist statistisch erfaßbar und numerisch

quantifizierbar und besitzt somit eine besondere Aussagekraft.

2.1.3. Modalität des Fremdsprachentestens

Nachdem ein Überblick über die Gebrauchsbedingungen geschaffen wurde, und die Gütekriterien und deren Anforderungen an Tests dargestellt wurden, ist es die Aufgabe dieses Kapitels, die Art und Weise wie die Fremdsprachenkompetenz erfaßt werden kann, darzustellen. Diese Darstellung soll nur insofern vollständig sein, als sie einen Bezugsrahmen für das Forschungsobjekt bilden soll. Hierzu ist es nützlich, einen Blick auf die geschichtliche Entwicklung zu werfen.

2.1.3.1. Geschichtliche Einführung

Als Teil der Fremdsprachenmethodik folgte das Testen den allgemeinen methodischen Entwicklungen.
Spolsky (1978) diskutiert drei größere "Trends" des Fremdsprachentestens, die er als vorwissenschaftlich (pre-scientific), psychometrisch-strukturalistisch (psychometric-structuralist) und integrativ-soziolinguistisch (integrative-sociolinguistic) bezeichnet. Diese drei Trends folgen aufeinander, überlappen sich aber zeitlich und auch methodisch.

2.1.3.2. Der vorwissenschaftliche Trend

Er ist gekennzeichnet durch die zentrale Stellung der beurteilenden Person, in der Regel einer Lehrperson, von der angenommen wird, daß sie die Leistung der Probanden beurteilen kann. Die statistischen Gütekriterien der Objektivität und Reliabilität sind dem subjektiven Urteil des Lehrers untergeordnet. Typische Techniken zur Leistungsfeststellung sind Übersetzung aus oder in die Fremdsprache, freier Aufsatz, Nacherzählung, mündliche Interviewtests und ausgewählte grammatikorientierte, literarische oder landeskundliche Aufgaben. Auf diese Tests wurde bereits bei der Besprechung der Gütekriterien hingewiesen. Sie erscheinen teilweise inhaltlich gültig, besitzen aber keine große Reliabilität. Dieser Trend, der praktisch seinen Anfang beim Beginn des Fremdsprachenunterrichts hat, ist nicht abgeschlossen. Subjektive Leistungse[illegible]ung wird auch weiterhin nicht

nur im Unterricht sondern auch in offiziellen Eignungstests und Abschlußprüfungen durchgeführt.

Wenngleich Probleme der Subjektivität teilweise durch verfeinerte Beurteilungsmethoden verringert werden können, bleiben zwei Probleme bestehen:

1) Es ist fraglich, inwieweit diese Techniken (besonders z. B. die Übersetzung) repräsentativ für die Erfassung einer allgemeinen Fremdsprachenkompetenz sind.
2) Sie erfordern einen sehr großen Aufwand.

2.1.3.3. Der strukturalistisch-psychometrische Trend

Dieser Trend entstand aus dem Zusammenwirken von strukturalistischer Linguistik und psychometrischer Testmethode. Auf der theoretischen Seite stimmen beide Richtungen in der behavioristischen Ansicht überein, daß Sprache ein auf 'habits' beruhender Prozeß ist. Auf der praktischen Seite benötigt die psychometrische Testtheorie analysierbare objektivierbare Einzelaufgaben, die von einer strukturalistischen Grammatik leicht geliefert werden können.

Der erste bedeutende Hauptvertreter ist Lado mit dem Werk "Language Testing" (1961). Auch die erwähnten Werke von Valette (1967), Harris (1969) und Clark (1972) folgen größtenteils dem strukturalistisch - psychometrischen Trend.

Carroll (1961) bezeichnet das daraus erwachsene Verfahren als 'discrete-point-approach'.

Davies (1968:53) sagt:

> "Whether one is testing competence in a native language or in a foreign language, it is useful to start from a classification of different areas of language competence in which individual differences are to be sought and measured."

Er präsentiert eine solche Klassifikation in der TAFEL 4.

TAFEL 4

		Phonology and Orthography	Lexicon: Morphemes, Words, Idioms	Lexicon: Semantic and Grammatical Components of Lexicon	Grammar: Morphology and Syntax	Grammar: Semantic Components
Spoken Language	Receptive Skills (Listening)	Phoneme recognition and discrimination; ability to discriminate words or phrases differing in one phoneme or distinctive feature [List phoneme pairs to be discriminated]	Recognition of lexical elements as belonging to the language [List lexicon]	Recognition of semantic and grammatical meanings (i.e. word class assignments) of lexical elements	Recognition of morphological and syntactical features and associated phonology [List]	Understanding
	Productive Skills (Speaking)	Phoneme production; ability to produce phonemes or allophonic variants in word forms or phrases, with accuracy at either phonemic or phonetic level [List phonemes of language with variants]	Ability to produce lexical elements fitting semantic and grammatical specifications		Ability to produce Morphological and syntactical features with appropriate phonology	In appropriate contexts
Written Language	Receptive Skills (Reading)	Recognition of the graphemic symbols of the language, with (as appropriate) ability to name them and give their sounds [List graphemes]	Above, plus recognition of meanings and pronunciation of written forms (including special graphemic symbols, abbreviations, etc.) [List]		Above, plus recognition of special grammar-related conventions of the written language, such as punctuation, capitalization, certain spelling changes, etc. [List such conventions]	
	Productive Skills (Writing)	Ability to write (by hand or other method) the graphemes of the language, state their customary ordering [List any details not covered above]	Spelling		As for receptive skills, plus ability to produce written conventions in appropriate contexts	

. Chart of linguistic competences to be filled out for a language system

(entnommen aus Davies, 1968:54,55)

Oller definiert den 'discrete-point-approach' näher:

"Traditionally, a discrete point test is one that attempts to focus attention on one point of grammar at a time. Each test item is aimed at one and only one element of a particular component of a grammar (or perhaps we should say hypothesized grammar), such as phonology, syntax, or vocabulary. Moreover, a discrete point test purports to assess only one skill at a time (e.g. listening, or speaking, or reading or writing) and only one aspect of a skill (e.g. productive versus receptive or oral versus visual). Within each skill, aspect, and component, discrete items supposed by focus on precicely one and only one phoneme, morpheme. lexical item, grammatical rule, or whatever the appropriate element may be." Oller (1979:137)

Arten von 'discrete-point tests' sind z.B. Grammatiktests, Wortschatztests,

phonetische Diskriminationsaufgaben. Man muß hier allerdings vorsichtig mit Generalisierungen sein, denn nicht alle Tests dieser Art sind notwendigerweise discrete-point-tests.

Im Deutschen wird gelegentlich der Ausdruck 'Formative Tests' gebraucht für Tests, die nur eine Fähigkeit messen. Dieser Ausdruck ist unscharf, da er sich auch auf subjektive Tests bezieht, die nur eine linguistische Teilfähigkeit messen.

Wegen seiner klaren Verständlichkeit und seiner allgemein erfolgten Zuordnung zur psychometrisch-strukturellen Richtung soll der Terminus 'discrete point test' beibehalten werden. Als Beispiel seien hier einige typische Aufgaben ausgewählt aus Lado (1961:57):

"A w o r d a n d t w o p i c t u r e s . The examiner says s h e e p and the students choose from the two pictures that follow:

Picture B fits the word given by the examiner. The student who cannot discriminate between the vowel of s h e e p and that of s h i p is at a loss in choosing his answer.

A s e n t e n c e a n d t w o p i c t u r e s . The examiner says H e i s w a t c h i n g t h e w i n d o w . The students choose from the following pictures:

Students who do not hear the difference between the middle consonants of w a t c h i n g and w a s h i n g are at a loss to select the right picture."

und von Harris (1969):

Auditory discrimination (Harris, 1969:32)
Der Schüler hört z.B. die Wörter
(1) (a) cot (b) caught (c) cot
Seine Aufgabe ist es, zu bestimmen, welches Wort sich von den anderen unterscheidet.

Grammatical structure (Harris, 1969:26-28)
(2) Mary _____ in New York since 1960.
(a) is living (b) has lived (c) lives

Sentence interpretation
(3) An old friend of John' s brought him news of his uncle last night.
Him refers to (a) an old friend (b) John (c) the uncle

Word order
(4) When _____ _____ _____ _____?
(a) plan (b) you (c) to go (d) do

Vocabulary (Harris, 1969:52)
(5) A brief, light sleep
(a) yawn (b) nap (c) struggle (d) hug

(6) The old woman was too _____ to push open the door.
(a) harsh (b) deaf (c) sincere (d) feeble

Diese Methode läßt folgende Vorteile erkennen:
Unter der Annahme, daß Sprachunterricht discrete-point-orientiert sein kann, liefern diese Tests eine diagnostische Information über Schwächen und Stärken der Probanden. Diese Tests sind leicht durchzuführen und auszuwerten. Sie sind objektiv und nach diesem Prinzip aufgebaute standardisierte Tests, z.B. der bekannte TOEFL (Test of English as a Foreign Language) der jährlich zweimal weltweit durchgeführt wird, erreichen nahezu perfekte Reliabilitätswerte.
Ohne auf inhaltliche Details einzugehen,seien einige Nachteile aufgezählt: Wie bei allen multiple-choice-Testverfahren ist ein beträchtlicher Aufwand zum Entwurf der Aufgaben, zum Vortesten, zur Revision und Reliabilisierung notwendig. In der Regel entsprechen die Aufgaben durch ihren discrete-point-

Charak. nicht dem normalen Sprachgebrauch. Oller sagt dazu:

> "It is rarely necessary for a student to understand whole sentences in order to answer discrete-point items correctly, and it is probably accurate to state that it is never necessary for a student to understand a context longer than a sentence in order to answer a discrete-point item. In fact, if it were necessary, the item would violate the cardinal principles of the discrete-point approach." Oller (1973:190)

Ein weiteres Problem besteht in der Auswahl repräsentativer Aufgaben. Welche 'skills' sollen in welcher Gewichtung getestet werden, um das Konstrukt einer Fremdsprachenkompetenz zu erfassen? Als Beispiel der Problematik seien hier zwei im deutschen Bereich vorhandene kommerzielle Tests vorgestellt: HET 6+, Fettweiß et al (1966) und Englisch Einstufungstest 6+, Schrand (1963).

Der HET 6+ besteht aus sechs Subtests: 1) Wortschatz; 2) Hörverständnis; 3) Rechtschreibung; 4) Leseverständnis; 5) Grammatik; 6) Aussprache.

Der Englisch Einstufungstest besteht aus den fünf Subtests: 1) Spelling; 2) Vocabulary; 3) Structure; 4) Pronunciation; 5) Listening Comprehension.

Ohne Angabe von Gründen werden jeweils verschiedene Teilkompetenzgebiete ausgewählt und gehen mit unterschiedlicher Gewichtung in die Gesamtwertung der 'Fremdsprachenleistung' ein. Diese Frage der Gewichtung stellt sich bei jedem discrete-point orientierten Test und kann als Grundproblem für die Validierung des unter einem strukturalistischen Aspekt gesehenen Konstrukts der Fremdsprachenkompetenz betrachtet werden.

Neben die Fragwürdigkeit, ob und auf welche Weise die Fremdsprachenkompetenz zu ihrer testtechnischen Erfassung in Einzelskills aufgeteilt werden kann, tritt also die Frage, ob discrete-point tests in der Lage sind, die Fähigkeit eines Probanden zu zeigen, ob er auch in tatsächlichen Kommunikationssituationen "funktionieren" kann.

Carroll (1961) stellt fest, daß die Beantwortung separater Aufgaben ohne weiteres möglich ist, ohne daß ein Proband eine tatsächliche Kommunikationsfähigkeit besitzt:

> "... if we limit ourselves to testing only one point at a time, more time is ordinarily allowed for reflection than would occur in a normal communication situation, no matter how rapidly the discrete-points are presented." (1961:318)

Auf dieser Grundlage beruht auch das sog. M o n i t o r M o d e l l von Steve Krashen (1977 a/b). Auch Krashen sagt, daß bei 'discreten' Einzelaufgaben der Proband seine Aufgabenbeantwortung überwacht (to monitor), was bei integrativen Tests nicht in dem Maße möglich ist.

Dazu sagt Perren (1967):

> "The constituent elements of language skills do not always add up to the total of ability in communication which is actually achieved. [He, therefore, argues] for deliberately using tests of gross skills of communication rather than concentrating exclusively on tests of their assumed constituent elements." Perren (1967:28)

Brière folgt:

> "... traditional linguistic definitions of the notion of competence in a language are too narrow and are inadequate in identifying all of the skills involved when two people communicate. Consequently, discrete item language tests based on a narrow definition of linguistic competence will be inadequate." Brière (1972:324)

2.1.3.4. Der integrativ-soziolinguistische Trend [10c]

Als Reaktion auf die Unzulänglichkeiten der discrete-point tests sind im Gefolge einer soziolinguistischen Neuorientierung des Fremdsprachenunterrichts integrative Tests, d.h. Tests, in denen mehrere 'skills' zusammen erfaßt werden, in eine neue Position gerückt.
Auch im Strukturalismus gibt es Tests, die integrativ sind, wie z.B. Hörverstehen, wobei Wortschatz, Grammatik, etc. gleichzeitig geprüft werden. Sie werden jedoch immer als separat erfaßbare Teilkompetenzen gesehen. Sie sind nicht integrativ in Bezug auf eine Messung der gesamten Fremdsprachenkompetenz. Im sog. 'integrative-sociolinguistic approach' wird das Zusammenwirken aller Skills als Bedingung zum Gelingen des Kommunikationsaktes gesehen.
Die Beherrschung von Teilkompetenzen ist dem Leitziel der Kommunikativen Kompetenz[11] untergeordnet. Eine Messung der Sprachkompetenz bedeutet daher nicht die Erfassung von Einzelelementen und ihre Addition, sondern eine Messung der integrativen Beherrschung der Elemente zum Zweck des Gelingens eines kommunikativen Prozesses. Der Test soll deshalb einem tatsächlichen Kommunikationsprozeß so ähnlich wie möglich sein.
Erbringen discrete-point Tests eine quantitative Kapazitätsbestimmung der Fremdsprachenleistung (wieviel beherrscht ein Pb von den Einzelementen einer Fremdsprache), so erwartet man von integrativen Tests eine funktionale Bestimmung qualitativer Art der Fremdsprachenleistung. Das Testergebnis soll aussagen, wie gut ein Proband seine Sprachkenntnisse in Kommunikationssituationen verwenden kann.

10c Nach Valette, 1977:308, 'psycholinguistic - sociolinguistic'
11 Zur Begriffsklärung siehe 2.2.

Diese unterschiede lassen eine deutliche Beziehung zur Testintentionalität erkennen (siehe 2.1.1.).

Davies drückt dies so aus:

> "Apparently the extent to which a language test should attempt to measure specific aspects of competence depends upon its purpose - that is the extent to which there is need for diagnosis of specific skills as opposed to a generalized, overal assessment of proficiency." Davies (1960:57,58)

Davies greift hier zurück auf das, was Carroll (1961) als 'disjunctive fallacy' bezeichnet, und was Ingram wie folgt beschreibt:

> "It is, in many cases, quite unnecessary to suppose that one has to make an either/or choice, that if one approves of integrative tests, one should therefore disapprove of discrete-point ones. This disjunctive 'fallacy', as Carroll calls it, stems, it seems to me, from misunderstanding about the nature of language command." Ingram (1978:12)

Diese Aussage findet auch einen statistischen Nachweis in der Untersuchung von Farhady (1979).

Bei einem integrativen Test scheint das Problem des 'sampling' zu verschwinden. So stellt Carroll fest:

> "Indeed this "integrative" approach has several advantages over the "discrete-structure-point" approach. It entails a more diffuse sampling over the total field of linguistic items and thus depends less upon the specifics of a particular course of training." Carrol (1965: 396-397)

Damit löst sich gleichzeitig das Problem der Konstruktvalidierung, denn der Testinhalt muß nicht validiert werden in Bezug auf ein Konstrukt, wie das beim discrete-point Test der Fall ist, wo das Testziel,die Sprachfähigkeit, nicht direkt gemessen wird, sondern wo verschiedene Teilaspekte des zu erfassenden Konstrukts gemessen werden.

Bei integrativen Tests sind Testinhalt und -ziel im Idealfall identisch. Man kann also von logischer Gültigkeit im Falle eines Testinterviews sprechen, wenn man annimmt, daß ein Interview - so wie es im Test geführt wird - eine repräsentative Stichprobe für einen tatsächlichen Kommunikationsprozeß ist. Diese Annahme muß aber erst bewiesen werden.

Oller hat dieses Problem durch die Einführung des Begriffs der 'pragmatischen Tests' angegangen, d.h. die Gültigkeit der Tests ist bereits durch ihre inhaltliche Bestimmung gegeben (Oller, 1979:36 ff).

Pragmatische Tests sind bestimmt dadurch, daß sie natürlichen Kommunikationssituationen entsprechen. Sie müssen die sog. Bedingungen der Natürlichkeit, 'naturalness criteria', erfüller:

> "Pragmatic language tests must meet two naturalness criteria: first they

> must require the learner to utilize normal contextual constraints on sequences in the language, and second, they must require comprehension (and possibly production also) of meaningful sequences of elements in the language in relation to extralinguistic contexts." Oller (1979:70)

Somit zeichnen sich pragmatische Tests aus als Untergruppe von integrativen Tests, die mehrere Sprachelemente einschließen, aber nicht notwendigerweise die Natürlichkeitsbedingungen erfüllen müssen. Durch diese Unterscheidung ist die bis dahin übliche Austauschbarkeit der Begriffe 'integrative Tests' und 'pragmatische Tests' nicht mehr möglich. (Vgl. Oller, 1979:38) Wesentlich ist, daß pragmatische Tests den extralinguistischen Kontext mit einschließen. Dies trennt sie klar von discrete-point tests, die dies völlig ausschließen[12].

Ob eine Testaufgabe die Bedingungen eines pragmatischen Tests erfüllt, kann empirisch, logisch (durch Definition) oder durch ein Experiment festgestellt werden, aber nicht durch subjektive Erhebungen.
Als Beispiele für pragmatische Tests führt Oller das Diktat und den CT an (siehe Kapitel 2.4.).

2.2. Die Fremdsprachenkompetenz

Bei der Darstellung der Testmodalität wurde ihre Interdependenz zur Sichtweise der Fremdsprachenkompetenz deutlich. Im folgenden Kapitel wird eine Darstellung des Begriffs der Fremdsprachenkompetenz in Relevanz zur vorliegenden Arbeit gegeben.

Drei für das Testen grundlegende Annahmen seien vorausgeschickt:

1) Es gibt unterschiedliche Grade der Fremdsprachenkompetenz.
2) Die Grade der Fremdsprachenkompetenz können durch formalen Unterricht gesteigert werden.
3) Die Fremdsprachenkompetenz ist im statistischen Sinne normal verteilt.

Der deutsche Begriff Fremdsprachenkompetenz wird hier für das englische 'Foreign Language Proficiency' gesetzt. Wenngleich der Begriff der Kompetenz durch verschiedene Interpretationen vorbelastet ist[13], ist er dem Begriff 'Fähigkeit' vorzuziehen, der durch seine Entsprechung zu 'skills' gebunden ist. Eine brauchbare Definition des Begriffs der Fremdsprachenkompetenz (Foreign

---

12 Siehe Lado (1961:26)

13 Siehe auch Hecht (1980:10)

Language Competence) gibt Brière:

> "[Foreign Language Competence is] the degree of competence or the capability in a given language demonstrated by an individual at a given point in time independent of a specific textbook, chapter in the book or pedagogical method." (Briere, 1972:322)

Der Begriff der Fremdsprachenkompetenz ist abzugrenzen gegen den Begriff der 'Fremdsprachenbegabung' (aptitude) und 'Fremdsprachenleistung'.

Die Fremdsprachenbegabung wird gesehen als die allgemeinen Voraussetzungen und Veranlagungen zum erfolgreichen Erlernen gleich welcher Fremdsprache (siehe Aptitude Test der Matrix "Testarten").

Die Fremdsprachenleistung ist das tatsächlich erbrachte Resultat einer Person in Bezug auf bestimmte Anforderungen und Zielsetzungen eines Fremdsprachenlehrangebots. Die Leistung kann als statisch oder dynamisch interpretierte Größe gesehen werden, wobei die letztere wegen der Vielzahl sehr schwer kontrollierbarer endogener und exogener Faktoren kaum greifbar gemacht werden kann. [14]

Ein tieferes Eindringen in den Kompetenzbegriff zeigt, daß er verschieden interpretiert werden kann, d.h. daß ihm verschiedene Modelle zugrunde liegen. Bei der Darstellung der Testmodalitäten wurde bereits andeutungsweise auf zwei Modelle hingewiesen:

Das erste ist das Modell der strukturalistischen Linguistik, das - von einer Oberflächenbetrachtung ausgehend - isolierbare Sprachelemente erkennt und annimmt, daß sich eine Fremdsprachenkompetenz, sofern es sie gibt, in diesen sprachlichen Fertigkeits- und Leistungsbereichen niederschlägt. Dabei ist jeder außersprachliche Bezug nicht legitim zur Darstellung der Komponenten siehe die Matrix von Davies, TAFEL 4).

Auch bei Chomsky, der das strukturalistische Modell wesentlich erweiterte, ist dieser Rahmen bestimmend. 'Competence' wird definiert als

> "the ability of the idealized speaker / hearer to associate sounds and meaning strictly in accordance with the rules of his language." Chomsky (1968:[illegible])

Das Modell besteht in der Beschreibung der Transformationsregeln, die eine Beziehung herstellen zwischen der Oberflächenrealisierung und der zugrundeliegenden Tiefenstruktur. Dabei ist die linguistische Kompetenz eine abstrakte Darstellung dessen, was ein idealer Sprecher/Hörer hypothetischerweise

---

14 Zur Problematik der Aspekte des Leistungsbegriffs siehe Gaude/Teschner, (1970) und Steltmann (1978).

mit seiner Kenntnis der Sprachregeln auf der Oberflächenstruktur der Sprache tun kann. Das Modell des idealisierten Sprechers/Hörers ist aber keine Beschreibung der Sprache im tatsächlichen Gebrauch. Dies wird dem Gebiet der Performanz überlassen und ist nicht Teil der Kompetenz. Dieses Modell entspricht dem 'discrete-point approach', ist aber unbrauchbar für pragmatische Tests, denn es fehlt die die soziale Perspektive einschließende Dimension der Sprache. Es ist nicht brauchbar für den Fremdsprachenunterricht.

Hier soll der Begriff der Kompetenz so verwendet werden, wie er von Dell Hymes und Cooper (1968) erweitert wurde[15]. Dieses Modell ist grundlegend für die gegenwärtige Fremdsprachendidaktik. Es besteht aus mehreren Schichten:

Ähnlich wie beim oben dargestellten strukturalistischen Modell werden linguistische Fähigkeiten und Fertigkeiten erkannt; dies sind die phonologische Teilkompetenz, die lexikalische Teilkompetenz, die grammatikalische Teilkompetenz und die semantische Teilkompetenz. Sie bilden zusammen die Schicht oder Ebene der 'Code Competence' und dienen der Übermittlung von 'signification'.[15a] Neben die Code Competence tritt die Kommunikative Kompetenz, die Fähigkeit zum sprachlichen Handeln.

In Anlehnung an Halliday (1975) und im Gegensatz zu Dell Hymes wird hier die Trennung von Kompetenz und Performanz nicht beibehalten, sondern "knowledge and ability for use" als Einheit im Sprecher angesehen.

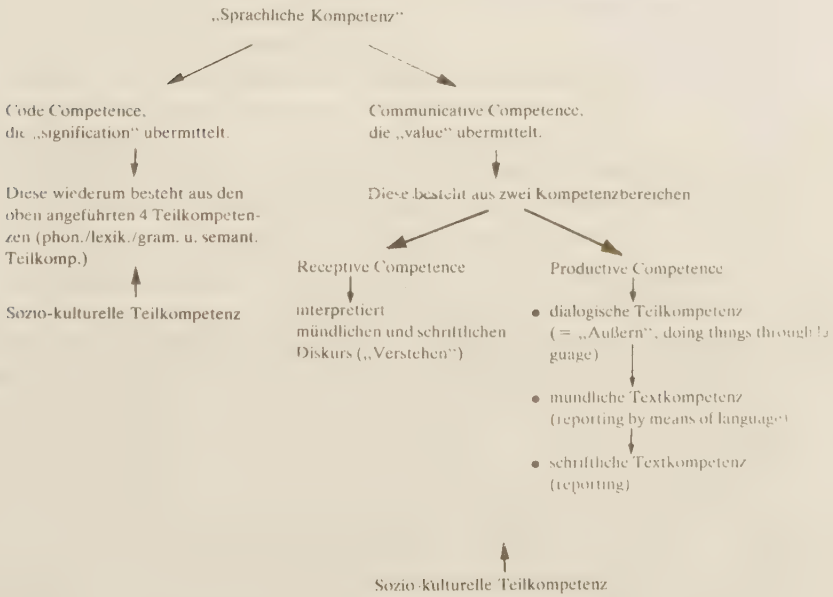
Nach Hecht befähigt dieses Kompetenzmodell den Sprecher zur

1) soziokulturellen Orientierung in der situativen Dimension (→ context appropriacy, cultural knowledge → expressing social meaning);
2) sozio-semantischen Orientierung in der linguistischen Dimension (er kann sich im Sinne der Wilkins'schen notionalen Kategorien semantisch-grammatikalisch richtig verhalten → er vermittelt konkrete Bedeutung → expressing propositional meaning, und er kann die Wahrheit einer Aussage modifizieren → expressing modal meaning); und
3) pragmatischen Orientierung in der intentionalen Dimension (er beherrscht den mündlichen und schriftlichen Diskurs, um mit der Sprache zu handeln oder durch sie zu berichten → expressing pragmatic meaning).

15 Siehe eine ausführliche Darstellung bei He t (1980:9-16)
15a Nach Hymes & Cooper, in: Hecht, 1980:9-16.

Dieses Modell deckt sich auch mit den Einzelleistungen, die vom Sprecher erbracht werden müssen bei der Realisierung des Sprechhandelns, d.h. im eigentlichen Sprechakt, bestehend aus lokutivem Akt, illokutivem Akt und perlokutivem Akt (dem Modell liegen die Kategorien von Austin, 1962, zugrunde). Für das linguistische Element im Sprechakt wird die Code Competence angesetzt, für die beiden anderen die Communicative Competence. Insgesamt ergibt sich unter dem Oberbegriff der sprachlichen Kompetenz folgendes Gesamtmodell:

ABB III



(entnommen - in Abänderung - aus: Hecht, 1980:14)

Die zusätzlich eingefügte soziokulturelle Teilkompetenz ist dadurch bedingt, daß Kommunikation nicht nur auf digitaler Ebene (verbale inhaltsbezogene Mitteilung) und analoger Ebene (Herstellung der interpersönlichen emotionalen Beziehung), sondern auch im Kontext eines bestimmten Landes und einer speziellen Sprachgemeinschaft stattfindet. Sie beeinflußt das Modell auf allen Stufen.

Zwei wesentliche Kennzeichen dieses Modells seien nochmals hervorgehoben:

1) Die 'Sprachliche Kompetenz' wird bestimmt durch das Zusammenwirken aller Teilkompetenzen, wie die zuletzt erwähnte soziokulturelle Teilkompetenz beispielhaft zeigt.

2) Das Modell ist in mehrfacher Hinsicht kontextbezogen.

Einen besonderen Aspekt der Fremdsprachenkompetenz hat Oller mit der sog. ' E x p e c t a n c y G r a m m a r ' entwickelt.[16] Sie beruht auf dem 'Analysis-by-Synthesis' - Konzept aus den 50er Jahren, das eine enge Verwandtschaft zwischen rezeptiven und produktiven Fähigkeiten plausibel erscheinen läßt. Grob vereinfacht sieht dieses Konzept so aus:

Sprachperzeption und -rezeption sind grundsätzlich mit der Fähigkeit des Sprachbenutzers verbunden, für jedes ankommende Signal eine interne Entsprechung zu erzeugen. Es wird dazu die Hypothese aufgestellt, daß der Sprachbenutzer ein ankommendes Signal analysiert, indem er es synthetisiert. Der Empfänger nimmt somit selbst aktiv am Verstehensprozeß teil. Eine Erweiterung dieser Hypothese sagt, daß auch alle produktiven Vorgänge in ähnlicher Weise zu beschreiben sind; sinnvolle Abfolgen von Elementen sind von einem Plan oder von einer Kommunikationsabsicht in der Art geleitet, daß das erzeugte Produkt (gesprochen oder geschrieben) bei seiner Freisetzung kontrolliert wird, ob es mit der tatsächlich beabsichtigten Bedeutung übereinstimmt.

Auf dieser Basis wird von Oller ein Kommunikationsprozeß folgendermaßen dargestellt:

> "A person speaking or writing is planning what to say next and monitoring the output to see whether or not it matches the intended

16 Da der Terminus noch keine deutsche Entsprechung hat aber leicht verständlich ist, soll er beibehalten werden. Siehe Oller (1976:165-175) und (1979:16-35) sowie (1972:151)

> meaning. A person listening or reading on the other hand is constantly generating hypotheses about what will come next in the sequence in terms of what the writer or speaker is intending to say. These hypotheses of the receiver are quite analogous to the plans of the sender. If the speaker's plans gibe closely with the hypotheses of the listener then communication is effective. If they fail to match well then communication breaks down." Oller (1976:67)

Dieses verbale Vorplanen des Sprechers und die hypothetisierenden Vorgriffe des Sprachempfängers auf syntagmatische Entwicklungen wird nach Oller durch die sog. 'Expectancy Grammar' gesteuert. Sie ermöglicht den Sprachbenutzern die schnelle Verarbeitung der auf sie eindringenden Sprachelemente und die angemessene Reaktion darauf (vgl. Carpenter/Jest, 1975). Dabei werden phonologische, semantische und syntaktische Prozesse in äußerst ökonomischer Weise durchlaufen und die ganzen sprachstrukturellen, inhaltslogischen und pragmatischen Erfahrungen des Sprachteilnehmers, die seine Erwartungen ausmachen, einbezogen. Das bezieht sich sowohl auf den rezeptiven als auch auf den produktiven Teil des Kommunikationsprozesses. Beide Vorgänge werden also als gleichermaßen kreativ angesehen.

Zusammenfassend wird die 'Expectancy Grammar' definiert als

> "a psychologically real system that sequentially orders linguistic elements in time and in relation to extralinguistic contexts in meaningful ways." Oller (1979:34)

Der Steuerungsmechanismus der 'Expectancy Grammar' wird als Grundlage aller - sowohl rezeptiver wie auch produktiver - Sprachfertigkeiten angesehen und steht somit an wichtiger Stelle eines Modells der Sprachkompetenz. Konsequenterweise müßte dies auch für niedrigere Niveaustufen einer Sprachkompetenz gelten.

Diese These ist zwar plausibel aber noch nicht bewiesen. Sang und Vollmer (1978:11) kritisieren diesen Ansatz. Sie argumentieren, daß Sprachproduktion und -rezeption nicht als Spiegelprozesse zu sehen sind.

Durch dieses Argument ist jedoch die These Ollers nicht widerlegt, denn wenngleich die Ausformung sprachlicher Prozesse unterschiedlich sein kann, so kann ihnen trotzdem eine gemeinsame Fähigkeit zugrunde liegen.[16a] Ollers Ansatz geht noch weiter: Die Existenz einer 'Expectancy Grammar' als Grundlage für jeden Sprachgebrauch bedingt einen allgemeinen Faktor der Sprachkompetenz, der unteilbar ist; Oller spricht von einem "General Language Proficiency Factor" und einer "Unitary Linguistic Competence". Nachweise für deren Existenz werden mit Hilfe faktorenanalytischer Verfahren gesucht. Diese Suche bringt plausible Ergebnisse (Oller, 1976: 170-173), die aber insbesondere aus statistisch-verfahrenstechnischen

---

16a Weitere Diskussionen und Forschungsergebnisse zu diesem Problemkreis cf. Aitchison 1976, Ervin-Tripp, Asher Naiman und Sticht diskutiert in: Hecht,1980:122ff, Lenneberg, 1967:99 und 393, Levin & Kaplan, 1971, und Woods, 1970:72.

Gründen von Sang / Vollmer (1978:32) kritisiert werden.
Jüngere Forschungen (Oller, 1979:423-458) Ollers untersuchen die Hypothese einer 'Unitary Language Competence' noch detaillierter und kommen zu einem positiven Ergebnis.
Auch Sang / Vollmer kommen in ihrer Untersuchung zu positiven, allerdings einschränkenden Ergebnissen.
Trotzdem dürfte die Diskussion, ob es einen allgemeinen Fremdsprachenfaktor gibt, und vor allem wie er interpretiert werden kann, noch nicht zu Ende sein. Für die vorliegende Untersuchung genügt es, wenn die 'Expectancy Grammar' einen Beitrag zur Erklärung des Cloze Phänomens liefern kann, damit das Experiment richtig durchgeführt werden kann.

Zusammenfassung:

Als mögliches Modell einer Fremdsprachenkompetenz wurde - abgrenzend gegen das die 'discrete-point Methode' bestimmende strukturalistische Kompetenzmodell - ein mehrschichtiges Kompetenzmodell vorgestellt. Es zeichnet sich aus durch die enge Verzahnung von Code Kompetenz und Kommunikativer Kompetenz mit einer Anzahl von integrierten Teilkompetenzen. Es ist in mehrfacher Hinsicht kontexteinschließend. Als besonderer Aspekt der Fremdsprachenkompetenz wurde das von Oller entwickelte Konstrukt einer sog. 'Expectancy Grammar' vorgestellt, das insbesondere zur Erklärung des Cloze Phänomens Verwendung finden kann.

2.3. Schulpädagogische und fachdidaktische Grundlagen

Die eingangs gestellte Forschungsfrage nach der schulspezifischen Anwendbarkeit des CT erfordert eine Betrachtung der schulspezifischen Hintergründe möglicher Strukturen der Verteilung der Kompetenz.

Im allgemeinen werden als Ausdruck der fremdsprachlichen Kompetenz von Schülern die durch die sechsstufige Notenskala ausgedrückten, von Lehrern erteilten Zensuren gesehen. Sie sind bei der Suche nach Verteilungsmustern der Fremdsprachenkompetenz unbrauchbar, da sie nicht an einem äußeren Kriterium orientiert sind, sondern sich jeweils nur auf die durch die jeweilige Lerngruppe bezogene Norm beziehen. Sie dienen primär zur interindividuellen Differenzierung der

Schüler . Außer der mit der Note verbundenen verbalen Wertangabe, z.B. "gut", die den Schüler im Hinblick auf das Lernziel seiner Lerngruppe charakterisiert, kann aus dieser Beurteilung kein objektiver Schluß, weder quantitativer noch qualitativer Art, auf seine Kompetenzstufe gemacht werden.

Erschwert werden Vergleiche dadurch, daß trotz existierender, teilweise minutiöser Ausführungsbestimmungen für die Zensurenvergabe keine qualitativen Bestimmungen vorhanden sind, die eine objektive Erfaßbarkeit und Vergleichbarkeit gestatten.[17] und [17a]

Auch Lehrpläne können kaum als Hilfe dienen. Selbst vereinheitlichte Lehrpläne können wegen der Bedeutung des interpretierenden Einflusses durch den personellen Vermittler (sprich 'Lehrer') keine objektiven Bewertungsgrundlagen liefern. Es mag zwar sein, daß Lehrer einer Schulart im Verhältnis zueinander eine gewisse Übereinstimmung erreichen können, das bedeutet aber immer noch einen hohen Anteil an Unzuverlässigkeit und beseitigt nicht die Problematik der Vergleichbarkeit verschiedener Kompetenzstufen der verschiedenen Jahrgangsklassen und der Unterschiede zwischen den Schularten.[17b] Bekannt sind die großen Divergenzen der Leistungsfeststellungen gerade im sprachlichen Bereich.

Es muß deshalb auf andere, objektive Beobachtungen zurückgegriffen werden, um Strukturen der Fremdsprachenkompetenzniveaus sichtbar zu machen. Verwieser sei in diesem Zusammenhang noch auf die bei der Behandlung der Fremdsprachenkompetenz aufgestellte Grundannahme, daß es unterschiedliche Grade der fremdsprachlichen Kompetenz gibt, und daß diese durch formalen Unterricht, wie er ein Grundprinzip aller drei Schularten ist, gesteigert werden können.

Ziel dieses fremdsprachlichen Unterrichts ist an den untersuchten Schularten und Jahrgangsstufen, die Sprache als Mittel zur Verständigung zu lehren.

Wollte man Unterschiede etwa in Bezug auf bestimmte Teilgebiete der Fremdsprachenkompetenz konstatieren, in der Art etwa, daß man sagt, gymnasialer Unterricht sei mehr auf eine schriftliche, der Hauptschulunterricht auf eine mündliche Fremdsprachenkompetenz hinorientiert, wäre das eine kaum belegbare Spekulation.

---

17 Siehe "Allgemeine Schulordnung und Ergänzende Bestimmungen", Donauwörth, (1979/80:§20.12.2.) und KMS, Nr. II/6-8/11002 vom 30. Januar 1980, "Bewertung von Leistungen in Englisch und Französisch an Gymnasien".

17a Selbst für die zentral gestellten Prüfungen des Abiturs, des Realschulabschlusses und des Qualifizierenden Hauptschulabschlusses liegen keine Angaben über ihre Gütekriterien vor, aus dem sich zuverlässige Aussagen ableiten ließen.

17b Zur ausführlichen Diskussion des Problemkreises siehe u.a. Steltmann, 1978, und Schröter, 1977.

Die Schularten

Die drei hier untersuchten Schularten sind Haupt-, Realschule und Gymnasium. Eine Vielzahl von Faktoren bestimmt die Zuordnung der Schüler in die jeweilige Schulart: Berufs- und Karrierewunsch, soziale Herkunft und soziale Zielvorstellungen von Eltern und Schülern, Selbsteinschätzung der Leistung der Schüler, die Beurteilung der Schülerleistung durch die Lehrer und Eltern und die Leistungsbereitschaft der Schüler. Diese Faktoren sind eng miteinander verbunden und üben individuell unterschiedlich starken Einfluß aus. Sie sind schwer objektiv zu bestimmen, was besonders für die motivationalen Faktoren gilt.
Demonstrierbar, wenngleich auch kaum objektiv quantifizierbar (siehe oben) ist die Bedeutung des Leistungsfaktors. Nur der Nachweis bestimmter Leistungen durch Schulzeugnis, Lehrerbeurteilung und Aufnahmeverfahren qualifiziert zur Aufnahme in ein Gymnasium.
Die Qualifikationsansprüche für Realschulen sind ähnlich, in der Regel aber nicht ganz so hoch.
Die Hauptschule als Pflichtschule stellt innerhalb der Norm keine Leistungsanforderungen.
Für den vorliegenden Zweck genügt es, diese zur Qualifikation nötige Leistung, die in den durch Zensuren ausgedrückten Beurteilungen der Fähigkeiten der Schüler in den Hauptfächern festgelegt ist, als allgemeine Schulleistung zu bezeichnen.
Die Fremdsprachenleistung hat keinen direkten Einfluß auf die Zuordnung der Schüler in die Schularten, es sei denn, ein Schüler wird wegen nicht erbrachter Leistungen im Fremdsprachenunterricht an eine niedrigere Schulart zurückverwiesen. Dieses negative, an einer allgemeinen Schulleistung orientierte Ausleseprinzip wird in der Praxis durch eine positive Auslese ergänzt, die vor allem bestimmt ist, durch ein allgemeines Bestreben leistungsfähiger Schüler, Realschulen und wenn möglich Gymnasien zu besuchen. Man kann deshalb ganz allgemein sagen, daß sich am Gymnasium die leistungsstärksten, an der Hauptschule die leistungsschwächsten und an der Realschule eine Mittelgruppe von Schülern befinden. Diese Aussage muß im Hinblick auf die Klassenstufen noch differenziert werden.
Die Zuordnung in die Schularten erfolgt vorwiegend zu bestimmten Zeitpunkten. Einmal tritt vor Beginn der Hauptschule, das ist im 5. Schuljahr, ein Teil der Hauptschüler in das Gymnasium über, ein weiterer, in der Regel kleinerer Teil folgt zu Beginn des 6. Schuljahres. Das heißt, daß noch ein Teil der potentiellen Gymnasiasten und alle Realschüler den 5. Schü-

lerjahrgang der Hauptschule besuchen. Nach dem 6. Schuljahr folgt in der Regel der Übertritt an die Realschule, d.h. daß ab der 7. Jahrgangsstufe ein Großteil der leistungsfähigeren Schüler die Hauptschule verlassen hat. Während in der Theorie auch danach noch eine Durchlässigkeit zwischen den Schularten besteht, d.h. daß Schüler von einer Schulart in die andere wechseln können, sieht es in der Praxis so aus, daß zwar Schüler, die bestimmte Ziele im Gymnasium oder in der Realschule nicht erreichen, zurückwechseln in die Realschule bzw. in die Hauptschule, aber kaum Schüler aufsteigen. Aus diesen Beobachtungen können einige Schlüsse gezogen werden.

1. Es besteht ein Leistungsgefälle vom Gymnasium zur Realschule und zur Hauptschule, das umso größer ist, je höher die Jahrgangsstufe.

2. Innerhalb der Schulart besteht ein durch die "Menge" des formalen Unterrichts geschaffenes Leistungsgefälle von den oberen zu den unteren Jahrgangsstufen.

3. Die einzelnen Jahrgangsstufen, besonders die einzelnen Lerngruppen, können wegen des jeweils gleichen Lernhintergrundes und wegen des Auswahlprinzips als relativ homogene Leistungsgruppen betrachtet werden.

Wurde bislang von der allgemeinen Schulleistung gesprochen, so soll ein Blick auf den Englischunterricht im besonderen geworfen werden: Das dargestellte Leistungsgefälle der allgemeinen Schulleistung kann auch im Bereich des Englischen erwartet werden.

1. Nach dem negativen Ausleseprinzip werden mit ungenügender fremdsprachlicher Kompetenz vom Gymnasium bzw. von der Realschule abgewiesen. Schüler, die Jahrgangsziele nicht erreicht haben, können nicht aufsteigen.

2. Die "Unterrichtsmenge" entspricht dem erwarteten Leistungsgefälle. Den meisten Englischunterricht erhalten Gymnasiasten, gefolgt von Real- und Hauptschülern. Detailliert sieht das so aus, nach Christ (1979:1-5)

WOCHENSTUNDENZAHLEN IM FACH ENGLISCH					
Schülerjahrgang	5	6	7	8	9
Hauptschüler	4	4	3	3	3
Realschüler	(4)	(4)	4	4	3
Gymnasiasten	6	6	4	4	3

TAFEL 5
(alle Angaben sind bezogen auf Bayern, Stand 1979)

Somit hat z.B. ein Hauptschüler nach dem 7. Schuljahr insgesamt erst 11 Wochenstunden Englischunterricht gegenüber 12 Wochenstunden des Realschü-

lers und 16 Wochenstunden des Gymnasiasten. Im 8. Schuljahr werden diese Unterschiede noch vergrößert.

3. Deutlich unterscheidet sich auch die fremdsprachliche Ausbildung der Lehrer der verschiedenen Schularten durch unterschiedliche Länge und Intensität des Studiums. Es soll ausdrücklich merkt werden, daß hier keine Kritik an den Leistungen verschiedener Lehrer oder Lehrergruppen geübt werden kann oder soll. Unterschiedliche Leistungen sind genauso individuell verschieden verteilt wie sämtliche andere Begabungen und Fähigkeiten. Sie sind objektiv kaum erfaßbar. Erfaßbar ist aber der unterschiedliche Ausbildungsgrad der Lehrer an verschiedenen Schulen.[18a] Lehrer an Gymnasien haben in der Regel ein mindestens achtsemestriges Fachstudium an einer Universität absolviert. Lehrer an Realschulen haben in der Regel entweder ein mindestens sechssemestriges Fachstudium an einer Universität oder eine fachliche Zusatzprüfung als Volksschullehrer absolviert. Lehrer an Hauptschulen haben in der Regel ein mindestens sechssemestriges allgemeines erziehungswissenschaftliches Studium mit einer Wahl-, Pflicht- oder Zusatzfachausbildung in der Didaktik des Englischunterrichts absolviert. Eine geringe Zahl von Lehrern haben eine Fachlehrerausbildung von einem Staatsinstitut oder gar keine spezielle Ausbildung. Aus diesen formalen Gründen können, wenn auch mit größter Vorsicht, Rückschlüsse auf die Unterrichtsqualität der verschiedenen Lehrergruppen gemacht werden, die sich auf die Fremdsprachenkompetenz der jeweiligen Schüler auswirken. Danach ist ein Qualitätsgefälle des Unterrichts zu erwarten, das dem dargestellten Muster der Fremdsprachenkompetenz entspricht oder zumindest nicht widerspricht.[18b]

Man könnte gegenbenenfalls erwarten, daß die Leistungsstruktur abhängig ist von den von den Lehrern verwendeten Vermittlungsmethoden, von den Lerninhalten und von den Lehrern selbst. Hierzu ist folgendes festzustellen: Obwohl bestimmte Richtlinien für die methodische Gestaltung des Unterrichts bestehen (z.B. Prinzip der weitgehenden Einsprachigkeit, induktive Grammatikvermittlungstechnik, Primat des Mündlichen), hat der Lehrer noch einen bedeutenden Spielraum bei der Unterrichtsgestaltung und beim Setzen von

---

18a Die Angaben beziehen sich auf Lehrkräfte, die nach der sog. alten Prüfungsordnung ausgebildet sind, d.h. auf alle in der Zeit der Durchführung des Experiments tätigen Lehrer.

18b Vorausgesetzt, die Ausbildungsgänge sind relevant für die Berufstätigkeit.

Schwerpunkten. Es gibt an allen Schularten sowohl progressive als auch traditionell orientierte Lehrkräfte. Es ist objektiv keine Struktur methodologischer Art nachzuweisen. Jegliche Aussage wäre rein spekulativer Art und bedürfte erst einer empirischen Untersuchung. Methodische Mischformen dürften eine solche Untersuchung komplizieren und klare Aussagen erschweren.

4. Der Stellenwert des Englischunterrichts liegt am Gymnasium und an der Realschule höher als an der Hauptschule. Dies drückt sich aus in höheren Stundenzahlen, einer zentraleren Stellung innerhalb der Bildungsziele und durch die Bedeutung der Englischleistung als Qualifikations- und Selektionskriterium an Gymnasium und Realschule. Der Selektionsvorgang hat außerdem noch einen häufig beobachtbaren motivationalen Einfluß auf die Hauptschüler. Sie fühlen sich nicht selten als Rest und Übriggebliebene aus diesem Selektionsvorgang, was ihre Leistungsbereitschaft dämpft und durch ein gesunkenes Selbstwertgefühl auch die Leistung beeinträchtigt. Besonders deutlich kommt dieses Phänomen in den B-Kursen der Hauptschule zum Ausdruck, die die letzte Stufe dieses Selektionsprozesses sind.

Z u s a m m e n f a s s u n g :

Die Betrachtung schulartenspezifischer Leistungsstrukturen läßt zwei Grundtatsachen erkennen, die auch für den Bereich des Englischen gelten. Sie können durch objektiv erfaßbare Beobachtungen belegt werden: das Schulsystem ist gekennzeichnet durch relativ homogene Lerngruppen. Es besteht ein Leistungsgefälle einmal zwischen oberen und unteren Jahrgangsklassen einer Schulart und zwischen den einzelnen Schularten. Dieses schulartbedingte Gefälle ist dadurch bestimmt, daß die Lerngruppen einen qualitativen Ausleseprozeß durchlaufen und daß sie verschiedene "Unterrichtsmengen" möglicherweise verschiedener Qualität erhalten.

Ob diese Situation wünschenswert ist oder nicht, sei dahingestellt; da es hier nur um eine deskriptive Erfassung der im Bereich der Untersuchung vorgefundenen Bedingungen zur Leistungsmessung geht, soll eine wertende Aussage nicht erfolgen.

2.4. Der Cloze Test

2.4.1. Entwicklung und Hintergrund

Das Cloze Verfahren wurde 1953 von Wilson Taylor eingeführt. Das Wort 'cloze' wurde abgeleitet von dem aus der Gestalttheorie kommenden Begriff der 'closure', womit die menschliche Tendenz bezeichnet wird, Lücken in Mustern auszufüllen. Der Test wurde konstruiert durch das Auslassen von Wörtern aus einem ausgewählten Text, die der Proband einzusetzen hatte. In seinem ersten Experiment schloß Taylor, daß das Cloze Verfahren eine effektive und zuverlässige Methode war, quantitativ und vergleichbar die relativen Schwierigkeitsgrade ('readabilities') verschiedener Stücke englischer Prosa für bestimmte Lesergruppen zu präsentieren (Taylor, 1953).

In der Folgezeit wurde eine Reihe von Versuchen unternommen und besonders zur konkurrenten Validierung des CT mit native speakers durchgeführt, die folgende drei Feststellungen im großen und ganzen zu bestätigen scheinen:[18c]

1) Das Cloze Verfahren ist ein relativ valides Meßinstrument des Leseverstehens.
2) Es ist ein relativ valides Meßinstrument des Schwierigkeitsgrades eines Textes.
3) Es ist ein relativ valides Instrument, die Angemessenheit des Schwierigkeitsgrades eines Textes für bestimmte Lerngruppen zu bestimmen.

Aus der Arbeit mit native speakers ist die Erforschung des CT mit non-native speakers im fremdsprachlichen Bereich erwachsen. Hier wurde das Verfahren als Meßinstrument im Hörverstehensbereich, als Instrument, bestimmte Reaktionen hervorzurufen ('Elicitation device'), als Meßinstrument zur Feststellung der Textschwierigkeit ('Readability'), als Lehrinstrument und als Meßinstrument der Fremdsprachenkompetenz untersucht und verwendet. Im Bereich der Fremdsprachenkompetenzmessung hat sich mittlerweile die sog. 'Standardmethode' eingebürgert, die Oller, der wohl die umfangreichste Forschung auf dem Gebiet geleistet hat, so beschreibt:

> "The standard test construction technique, known as the fixed ratio method, involves deleting every n^{th} word (where n usually varies from 5 to 10) and replacing each one with a standardized blank (usually about 15 typed spaces).
> The standard length of the cloze test is 50 items - thus the passage must be approximately 50 times n." Oller (1979:375)

Als Beispiele eines CTs seien zweier solcher Tests angefügt, die die Verfahrensweise veranschaulichen. Das erste Beispiel stammt aus Hinofotis, 1976:140 ff, das zweite aus Brown, 1979: 120.

18c Hier sei auf den kritischen Beitrag v[illegible] ander Adkins verwiesen, "Cloze tests and extensive reading tes[illegible] : MET, Vol.7, No.2, Nov.79.

(entno aus:Hinofotis, 1976:140 ff)

NAME_____ NATIVE LANGUAGE_____

Directions:

1. Read the passage quickly to get the general meaning.
2. Write one word in each blank.
3. Check your answers.

Note: Spelling will not count against you as long as the scorer can read the word.

Speed and Comfort

People traveling long distances frequently have to decide whether they would prefer to go by land, sea, or air.

Hardly anyone can enjoy sitting in a train for more than a few hours. Train compartments soon get cramped and stuffy. It is almost impossible to take your mind off the trip. Reading is only a partial solution, (1) for the monotonous rhythm of the wheels (2) clicking on the rails soon lulls you (3) to sleep. During the day, sleep comes (4) in snatches. At night, when you really (5) want to go to sleep, you rarely (6) manage to do so. If you are (7) lucky enough to get a sleeping compartment, (8) you spend half the night staring at (9) the small dim light in the ceiling (10) or fumbling to find your passport when (11) you cross a border if you are (12) traveling abroad. Inevitably you arrive at your (13) destination almost exhausted.

Long car trips are (14) even less pleasant, for it is impossible (15) to even read. But on highways you (16) can, at least, travel fairly safely at (17) high speeds.

By comparison, trips by sea (18) offer a great variety of comforts not (19) available on other types of transportation. You

(20) can stretch your legs on the spacious (21) decks, play games, swim, meet interesting people (22) and enjoy good food - always assuming, of (23) course, that the sea is calm. If (24) it is not, and you are prone (25) to sea-sickness, no form of transportation could (26) be worse. Even if you travel in (27) good weather, traveling by sea takes a (28) long time. Relatively few people are prepared (29) to sacrifice up to a third of (30) their vacation for the pleasure of traveling (31) on a ship.

Airplanes used to have (32) the reputation of being dangerous and some (33) experienced travelers are still intimidated by them. (34) They also have the disadvantage of being (35) the most expensive form of transportation. But (36) nothing can match them for speed and (37) comfort. You do not have to devise (38) ways of taking your mind off the (39) trip, for an airplane gets you to (40) your destination rapidly. For a few hours (41) you settle back comfortably to enjoy the (42) flight. The real escapist can watch a (43) good film and sip champagne on some (44) flights. But even when such refinements are (45) not available, there is plenty to keep (46) you occupied. Traveling at a height of (47) several thousand feet, far above the clouds, (48) at over five hundred miles an hour (49) is an exhilarating experience. An airplane offers (50) you an unusual and breathtaking view of the world, and the journey is so smooth that there is nothing to prevent you from reading or sleeping.

Acceptable-word Responses

Speed and Comfort
Every 7th word deleted

1. for, because
2. clicking, rolling, running
3. to
4. in
5. want, have, wish, need, plan
6. manage, get, want
7. lucky, rich, fortunate
8. you
9. the, a
10. or, and
11. you
12. traveling, going, still, from
13. destination, station, stop
14. even, much, certainly, always
15. to
16. can
17. high, normal, comfortable, fast, permitted
18. offer, provide, have, supply
19. available, found, provided
20. can, may
21. decks
22. and, even
23. course
24. it
25. to
26. be
27. good, fine, perfect
28. long
29. to
30. their
31. on
32. the, a
33. experienced, frightened, sensitive, fearful, few, brave, nervous
34. they, airplanes
35. the
36. nothing
37. comfort, efficiency
38. ways, means
39. trip, journey, time, speed
40. your
41. you
42. flight, trip, sight
43. good, nice, movie, fine, great, interesting
44. flights, airplanes
45. not
46. you, one, passengers
47. several, many, thirty
48. at, and, moving, flying
49. is
50. you, passengers, one

(entnommen aus: Brown, 1979:120)

Name _____
(Last) (First)

Birthdate: Day____ Month_____ 19__

Native language: _____

DIRECTIONS:

1. Read the passage quickly to get the general meaning.
2. Write only one word in each blank on the right.
   Contractions ("don't") and possessives ("John's" book) are one word.
3. Spelling will not count against you.
4. Check your answers when you are finished.

EXAMPLE: The boy walked up the street. He stepped on a piece of ice.
He fell (Ex.) but he didn't hurt himself. Ex. down

THE THINGS WE KNOW ABOUT LANGUAGE

Many things about languages are a mystery, and many will remain so. But some things we do know. First, we know that all human beings have a language of some sort. There is no race of men (1) on earth so backward that it (2) no language, or set of sounds (3) which the people can communicate with (4) another. In addition, in historical times, (5) has never been a race of (6) without a language. Second, there is (7) such thing as a primitive language. (8) are many people with undeveloped cultures, (9) are often considered uncivilized, but the (10) they speak are not primitive. In (11) known languages we can see complexities (12) must have been tens of thousands (13) years in developing. A third thing (14) know about language is the fact (15) all languages are perfectly adequate. That (16) , each is a perfect means of (17) the culture of the people who (18) the language. For example, French is (19) excellent language for expressing French culture, (20) it might be an inadequate language (21) expressing any other culture. Finally, we (22) that language changes. It is quite (23) for language to change. Change goes (24) in all aspects of language. Grammatical (25) change as do speech sounds. Changes in vocabulary are sometimes very extensive and may occur very rapidly. That is why vocabulary is the least stable part of a language.

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____
13. _____
14. _____
15. _____
16. _____
17. _____
18. _____
19. _____
20. _____
21. _____
22. _____
23. _____
24. _____
25. _____

ACCEPTABLE \NSWERS
THE THINGS WE KNOW ABOUT LANGUAGE

1. anywhere existing found here known living now
2. has knows
3. by through with
4. one
5. there
6. men people
7. no
8. there
9. that they which who
10. languages words
11. all many some
12. that which
13. of
14. to we
15. that
16. is means
17. communicating expressing transmitting
18. know speak(s) use(s)
19. an
20. although but however though yet
21. for in
22. know note see
23. common easy natural necessary normal usual
24. forth gradually on
25. aspects constructions features forms patterns rules structures

Diese Methode hat eine weite Annahme gefunden und wird insbesondere in den USA aber auch weltweit verwendet (siehe z.B. Alderson, 1978, in Algerien und Großbritannien; Anderson, 1972, in Papua, Neu Guinea; Anderson 1976 in Australien; Evans/Haastrup, 1976, in Skandinavien; Gefen, 1979, in Israel; Irvine & Oller, 1974, in Persien; Kržyzanowsky, 1976, in Polen; Lapkin & Swain, 1977 in Canada; Pike, 1973, in Japan, Chile und Peru; Stubbs & Tucker in Beirut; Waiman, 1979, in Israel; Wijnstra & Wageningen, 1974, in Holland[19]).

Mehrere Gründe machen das Verfahren besonders attraktiv: Es ist ein o b j e k t i v e s T e s t v e r f a h r e n , das relativ leicht zu konstruieren und durchzuführen ist. Der aufwendigste Teil scheint dabei die Textauswahl zu sein. Die Aufgabenstaltung erfolgt systematisch: nachdem gewöhnlicherweise einige Zeilen des Textes zur Orientierung der Probanden gegeben werden, wird jedes n-te Wort ausgelassen, bis die gewünschte Menge an Aufgaben erreicht ist. Dies ist beträchtlich einfacher als der Entwurf eines traditionellen objektiven Tests etwa mit Mehrfachwahlaufgaben. Rankin sagt hierzu:

> "The striking utility of the cloze procedure lies in the ease with which it can be used in test construction. This fact can best be appreciated if one compares the time and effort expended in writing a set of objective questions which have suitable reliability with the corresponding time and effort involved in simply deleting, say every 5th word in an article or some portion of the article." Rankin (1970:251)

Für die Durchführung empfiehlt Oller als groben Richtwert etwa eine Minute Arbeitszeit pro Item für Erwachsene, das sind bei 50 Items 50 Minuten, und etwas mehr bei jüngeren Testteilnehmern.Diese Länge ist durchaus akzeptabel und praktikabel, wenn man vergleicht, daß standardisierte objektive Tests in der Regel mehrere Stunden beanspruchen. Der HET 6+ von Fettweiß et al. (1966) benötigt 135 Minuten verteilt auf mehrere Tage, der Englisch Einstufungstest 6+ von Schrand et al. (1973) benötigt fünf Schulstunden an verschiedenen Tagen.

Das Verfahren besitzt eine große p s y c h o l o g i s c h e G ü l t i g k e i t ('face validity'). Der Proband kann ohne weiteres erkennen, daß der CT auch einer Erscheinung des täglichen Sprachgebrauchs entspricht, bei dem fehlende sprachliche Elemente ersetzt werden müssen. Clark (1978) argumentiert, daß diese Situation relativ selten ist, und daß der CT deshalb nicht voll dem normalen auch der Fremdsprache entspricht. Er bezeichnet deshalb den CT als indirektes Verfahren zur Mes-

19 Eine ausführliche Übersicht bieten die Bibliographien von: Jongma, 1971; Robinson, 1972; Riley, 1973; Oller, 1975; Boyce, 1976.

sung der Fremdsprachenkompetenz.

Dem ist entgegenzuhalten:

1) Auch wenn im täglichen Sprachgebrauch Texte nur selten zerstört sind, ändert dies nichts an der Tatsache, daß eine Lücke eine normale Erscheinung in einem Kommunikationsprozeß ist. Hierzu gehören z.B. akustische Störungen bei der Lautübertragung durch Nebengeräusche o.ä. und Zerstörungen eines geschriebenen Textes durch Risse, Flecken u.ä.

2) Lücken können nicht nur durch eine von außen kommende Zerstörung entstehen, was seltener vorkommt; sehr häufig entsteht im Kommunikationsprozeß mit der Fremdsprache das, was man als 'innere Lücke' bezeichnen kann: Jedes unbekannte Wort ist für den Lernenden eine Lücke, sei sie nun repräsentiert durch eine unverständliche Kombination von Lauten oder Schriftzeichen, wie das normalerweise vorkommt, oder sei es ein freier "Fleck", wie das beim Cloze Verfahren der Fall ist.

3) Oller argumentiert, daß der Vorgang des CTs genau dem natürlichen Vorgang der sog. 'Expectancy Grammar' entspricht (1971 passim; 1976:165-174; 1979:154 ff). Sie ist der grundlegende Prozeß eines jeden Kommunikationsprozesses. Dabei stellt ein Signalempfänger im Kommunikationsprozeß konstant Hypothesen auf in Bezug auf die zu erwartenden Signale. Er nützt hierzu, wie es Oller bezeichnet,

> "contextual constraints on sequences in the language, and second ... comprehension (and possibly production also) of meaningful sequences of elements in the language in relation to extralinguistic contexts." Oller (1979:70)

Damit entspricht der CT nicht nur auf der Ebene der physischen Erscheinung, sondern sozusagen auf einer tieferen, der psycholinguistischen Ebene, einer natürlichen Kommunikationssituation. CTs gehören somit zur Gruppe der sog. pragmatischen Tests, die eine Untergruppe der integrativen Tests bilden (siehe Oller, 1979:123 ff).

Auch in anderer Hinsicht ist der CT ein "direktes" Verfahren: Im traditionellen Verfahren muß der Proband häufig eine Aufgabe zu einem Problem lösen oder eine Antwort zu einer Frage geben. Er muß z.B. einen Ausdruck paraphrasieren oder eine Frage zu einem Text beantworten. Unter Umständen wüßte er zwar die Paraphrase per se, aber er kennt den zu paraphrasierenden Ausdruck gar nicht, oder er hat zwar den Text verstanden aber nicht die Frage. Dieses Problem der Aufgaben- bzw. Fragenstellung entfällt beim CT. Der CT verlangt und mißt direkt eine Antwort und nicht die Beantwortung einer Frage oder Aufgabe.

Somit entfällt auch der Gebrauch einer Metasprache, die mit dem zu testenden Gegenstand unter Umständen nichts zu tun hat (siehe z.B. HET 6+, Teil 4).

Eine weitere Vereinfachung, die der CT mit sich bringt, ist der W e g f a l l d e s ' s a m p l i n g ' . Während bei der traditionellen Kompetenzmessung abgewägt werden muß, welche 'Skills' und 'Subskills' in welchen Gewichtungen gemessen werden, nimmt man an, daß durch das systematisierte Aufgabenauswahlverfahren ein repräsentativer Querschnitt durch einen Text erfaßt wird.. Carroll sagt dazu allgemein für integrative Tests

> "Indeed this "integrative" approach has several advantages over the "discrete-structure-point" approach. It entails a more diffuse sampling over the total field of linguistic items and thus depends less upon the specifics of a particular course of training." Carroll (1965:396-397)

Dies trifft für den CT natürlich in besonderem Maße zu.

Als erste untersuchten Carroll, Carton und Wilds (1959) die Möglichkeit, den CT zur Feststellung der Fremdsprachenleistung zu verwenden. Ihre Zielgruppe waren High School -Schüler. Sie lehnten aber das Verfahren wegen relativ irreliabler Ergebnisse und zu vieler unkontrollierbarer Variablen ab. Seitdem wurden zahlreiche Studien unterschiedlicher Größenordnung mit verschiedenen Forschungsintentionen veröffentlicht, von denen sich viele auf oberflächliche Beobachtungen beschränken und wesentlicher qualitativer und quantitativer Angaben entbehren. Hier folgt ein Überblick über die verschiedenen wichtigen Studien und deren Ergebnisse:

2.4.2. Forschung zum Standard Cloze Test als Meßinstrument der Fremdsprachenkompetenz

1968 führte Darnell (1970) eine Untersuchung an 48 Studenten der Universität von Colorado durch. Wesentlicher Aspekt seiner Untersuchung ist die Auswertungsmethode, die er als 'clozentropy' bezeichnete, nach der von Shannon und Weaver (1949) entwickelten Informationstheorie. Shannon und Weaver entwickelten in "The Mathematical Theory of Communication" (1949) einen numerischen Ausdruck für die tatsächliche Informationsmenge (Entropie) einer Informationsquelle[19a] (H)

$$H = -\sum p_i \log_2 p_i \quad \text{oder} \quad -(p_1 \log_2 p_1 + p_2 \log_2 + p_n \log_2 p_n)$$

Dabei ist n die Anzahl unabhängiger Symbole, von denen jedes die Wahrscheinlichkeit p des Vorkommens besitzt (eine Beschreibung des Verfahrens siehe auch bei Reilly, 1971:350-353).

19a Von Shannon/Weaver, 1963:103, als 'measure of the degree of randomness' bezeichnet

Eine CT-Batterie aus Einzeltexten mit einer Lückenfrequenz, bei der jedes zehnte Wort ausgelassen ist, wurde von N= 48 Studenten bearbeitet. Das Besondere am Clozentropy Verfahren besteht darin, daß die Antworten nicht mit falsch oder richtig bewertet werden, sondern mit einer Gewichtung, die der Häufigkeit des Vorkommens der jeweiligen Antwort bei native speakers entspricht. Die Anzahl der Aufgaben/Lücken dieser Studie betrug n= 200. Die nach Hoyt berechnete Reliabilität dieses Tests betrug 0,86. Die Validierung der Clozentropy-Ergebnisse mit dem TOEFL (Test of English as a Foreign Language) ergibt einen Gesamtvaliditätskoeffizienten von 0,84, der besagt, daß ca. 70% der Varianz der Ergebnisse beiden Tests gemeinsam ist. Darnell weist darauf hin

> "Considering the structural and theoretical differences in the two tests, larger correlations could not be expected." Darnell (1968:40)

Darnell folgert, daß - nachdem der TOEFL-Test ein akzeptables Meßinstrument der Fremdsprachenkompetenz ist- die Cloze Battery ebenso akzeptabel sein muß. Es wird darauf hingewiesen, daß verschiedene CTs der Batterie unterschiedlich mit dem TOEFL (0,63 - 0,77) und seinen Subtests (0,49 - 0,70) korrelieren. Interessanterweise besteht die niedrigste Teilkorrelation zwischen der Gesamtclozebatterie und dem Leseverstehenssubtest (0,60). Die Ergebnisse im Bereich der Validität und Reliabilität sind zwar durchaus positiv, aber die komplizierte Auswertungsmethode macht das Verfahren sehr unpraktikabel.

Die erste Studie im Elementarschulbereich stammt von Anderson (1970). Er untersuchte "Primary school learners of English as a foreign language" in Papua, Neu Guinea. Keine Aussage ist gemacht über das Fremdsprachenkompetenzniveau dieser Gruppe. Sein CT bestand aus neun Passagen mit insgesamt 50 Aufgaben, die allerdings nach ihren Antwortmöglichkeiten bewußt ausgewählt wurden. Die Einzeltests erreichten Reliabilitätskoeffizienten von 0,8 und darüber. Der Gesamtreliabilitätsfaktor betrug 0,95. Diese Tests wurden allerdings nicht als 'Proficiency Tests' validiert, sondern mit dem für native speakers entworfenen Watt's Reading Comprehension Test und einem selbstentworfenen Leseverstehenstest. Der Korrelationskoeffizient betrug 0,78 für den Watt's Reading Comprehension Test und 0,85 für den selbstentworfenen Leseverstehenstest.
Zusätzlich wurde eine Faktorenanalyse durchgeführt, bei der sich nur ein gemeinsamer Faktor zeigte, der 65% der Varianz beinhaltete. Der CT hatte die größte Ladung dieses Faktors, der nach Anderson nur als Leseverstehen beschrieben werden konnte. Er schloß deshalb, daß der CT ein reliables und gültiges Meßinstrument des Leseverstehens ist.

Es soll auf zwei Eigenheiten der Studie Andersons hingewiesen werden, die seine Ergebnisse nicht direkt vergleichbar machen mit der vorliegenden Studie.

1) Das verwendete Verfahren mit der bewußten Aufgabenauswahl entspricht nicht dem Standardverfahren.

2) Der Test ist validiert als Meßinstrument des Leseverstehens. Geht man von einem strukturalistischen discrete-point-Standpunkt aus, kann von dem Meßergebnis des Leseverstehens nicht direkt auf eine allgemeine Fremdsprachenkompetenz geschlossen werden, denn auch die Erfassung anderer Teilkompetenzen wäre hierzu nötig.

Dem Konzept einer 'discriminant validity' von Campbell und Fiske (nach Stevenson, 1978) folgend, würde die Korrelation mit dem Test als Leseverstehenstest sogar die Erfassung einer allgemeinen Fremdsprachenkompetenz invalidieren. Campbell und Fiske (1959:81) konstatieren:

> "Tests can be invalidated by too high correlations with other tests from which they were intended to differ."
> Campbell und Fishe (1959:81)

Folgt man jedoch dem dargestellten mehrschichtigen Kompetenzmodell (siehe Kap. 2) und der Argumentation der sog. integrativen Testmethode mit (siehe K.2.1.3.4.) der Annahme, daß alle Teilkompetenzen eng miteinander verzahnt sind, kann geschlossen werden, daß CTs neben einer allgemeinen Fremdsprachenkompetenz überlappenderweise auch den Teilbereich des Leseverstehens messen, zumal der aufgezeigte Korrelationskoeffizient von 0,78 nicht allzu hoch ist.

In der Tat zeigt die folgende Untersuchung von Oller/Conrad einen sehr ähnlichen Korrelationskoeffizienten von 0,80 zwischen dem CT und dem Lesesubtest. Dies steht allerdings in gewissem Widerspruch zu dem bei Darnell gefundenen Wert von 0,60 für die gleiche Korrelation.

Im ersten Quartal 1970 testeten Oller und Conrad (1971) 102 ausländische Studenten aus 38 verschiedenen Ländern mit 27 verschiedenen Muttersprachen an der UCLA[20]. Die Probanden wurden in fünf Leistungsgruppen unterteilt, wozu noch zwei Vergleichsgruppen von native speakers (Freshman Composition und TESL Graduates) kamen. Der verwendete CT entsprach nach dem "Flesch Readability Index" dem College Niveau (Freshman and Sophomorres). Der Lückenabstand betrug 7 Worte für insgesamt 50 Aufgaben. Als richtig wurden nur die im Originaltext vorhandenen Wörter gezählt (von jetzt ab als 'exact scoring' bezeichnet). Die CT-Ergebnisse ('Mean

---

20 University of California at Los Angeles

Scores') entsprachen den "rankorders" der Leistungsstufen der Gruppen.

Mean scores of subjects on cloze test

	Group		N	Mean Score	Standard Deviation	Range
Non-Native Subjects	I	Beginning ESL	18	7.00	3.88	1-15
	II	Intermediate ESL	26	11.42	4.27	6-18
	III	Advanced ESL	26	18.38	4.54	14-30
	IV	Advanced Composition	23	21.00	4.11	14-28
	V	Graduates in TESL	9	20.92	4.68	12-29
Native Speakers	VI	Freshman Composition	26	21.38	4.04	16-28
	VII	Graduates in TESL	14	32.50	3.71	27-38

TAFEL 6
(entnommen aus: Oller/Conrad, 1971:189)

Gegen die Erwartung der Autoren wurden die Gruppen IV - VI jedoch sehr ähnlich eingestuft. Sie hatten erwartet, daß die Gruppen VI und VII als native speakers homogener in Bezug auf 'language skills' sind, und daß ein größerer Unterschied zwischen den non-native speakers, Gruppe IV und V einerseits und der native-Gruppe VI andererseits bestünde, wobei einige der non-native speakers die native speakers beträchtlich übertrafen. In Kenntnis der Verhältnisse an der UCLA und unter dem Gesichtspunkt der oben gegebenen Definition einer Fremdsprachenkompetenz, die nicht nur aus linguistischen Teilkompetenzen besteht, sondern auch kommunikative und soziokulturelle Teilkompetenzen einbezieht, ist dieses Ergebnis durchaus erklärlich:

Graduates und AdvancedComposition-Studenten haben in der Regel auch als Ausländer einen längeren Ausleseprozeß hinter sich, in den ein 'Freshman' erst eintritt. Kann man von einem native speaker gegenüber einem non-native speaker zwar u.U. eine höhere Codekompetenz erwarten, so können aber die kommunikative und soziokulturelle Teilkompetenz bei Ausländern, die - wie vor allem Graduates und fortgeschrittene Studenten in der Regel schon länger im Lande sind, durchaus weiter entwickelt sein.(Vgl. Alderson, 1980). Man kann schließen, daß diese über die linguistische Teilkompetenz hinausreichenden Fähigkeiten einen wichtigen Einfluß auf das Clozeergebnis haben.

Die Cloze Tests wurden zur konkurrenten Validierung mit dem UCLA ESLPE (English as a Second Language Placement Examination) korreliert. Der Gesamtkorrelationskoeffizient beträgt 0,88. Die höchsten Korrelationen des CTs mit Untertests des ESLPE wurde mit dem Diktat (0,82) und dem Leseverstehenstest (0,80) festgestellt. Wesentlich niedrigere Korrelationen bestanden mit dem Vokabeltest (0,59), dem Grammatiktest (0,58) und dem

Artikeltest (0,33).

Eine andere Studie von Oller (1972) war noch umfangreicher und hatte das Ziel, die Einflüsse von Auswertungsmethoden und Schwierigkeitsgraden des CT-Verfahrens bei der Messung der Fremdsprachenkompetenz zu untersuchen. Die CT-Ergebnisse wurden auch korreliert mit den Ergebnissen des UCLA ESLPE (Form 2A) und seinen Subtests als Kriterien externer Validität. Drei CTs unterschiedlicher Schwierigkeitsgrade wurden für die Studie entworfen. Jeder Text hatte 50 Lücken im Abstand von 7 Wörtern. Die 398 Teilnehmer waren Studenten an UCLA mit insgesamt 55 verschiedenen Muttersprachen, 132 von ihnen bearbeiteten den leichtesten CT I, 134 den mittleren CT II und 132 den CT III.
Als Ergebnis zeigte sich, daß die 'any-acceptable-word'- Methode der Bewertung bessere Reliabilitätskoeffizienten (0,90 - 0,95) und bessere Validitätskoeffizienten (0,80 - 0,89) als die 'exact-scoring'- Methode erbrachte und darüberhinaus die Probanden besser differenzieren konnte. Bei dieser Bewertungsmethode wurde die höchste Teilkorrelation (0,76 - 0,82) mit dem Diktat erreicht, gefolgt von Leseverständnistest (0,71 - 0,82), Grammatiktest (0,73 - 0,80) und Vokabeltest (0,63 - 0,75). Oller erklärte diese Ergebnisse damit, daß Diktat und Leseverständnis ähnlich den CTs integrativer sind als Grammatik- und Vokabeltests. Insgesamt wird die 'any-acceptable-word'- Bewertungsmethode als die günstigste betrachtet, und der CT wird als Test bezeichnet, der 'high-level integrative skills' erfordert. (Vgl. Alderson, 1978:371 und 1979:225)

Irvine, Atai und Oller (1974) untersuchten 159 persische Studenten an der Universität in Teheran, die sich einem CT mit 50 Aufgaben mit einem Lückenabstand von 7 Wörtern, zwei Diktaten und dem TOEFL unterzogen. Die CT-Ergebnisse lieferten einen Korrelationskoeffizienten mit dem TOEFL von 0,79 bei der Bewertung nach der akzeptablen-Antwort-Methode und von 0,78 bei der 'exact-scoring'-methode. Der CT korrelierte mit den kombinierten Diktatergebnissen mit 0,75 ('acceptable scoring') und 0,69 ('exact scoring'). Die Subtestkorrelationen sind der folgenden Tabelle zu entnehmen.

TAFEL 7

Pearson product moment correlation matrix for TOEFL part scores, total score, the cloze test scored by exact word and acceptable word methods, and two dictations

	List. Comp.	Engl. Struct.	Vocab.	Read. Comp.	Writ. Abil.	TOEFL Total	Accept. Cloze	Dict. Total	Exact Cloze	Easy Dict.	Hard Dict.
List. Comp.	1.0	.69	.56	.63	.68	.88	.76	69	.76	.70	.67
Engl. Struct.		1.0	.64	.57	.65	.85	.68	.63	66	.63	.63
Vocab.			1.0	.49	.60	.79	.51	.47	.49	.47	.47
Read. Comp.				1.0	.58	.77	.65	.53	.67	.53	.56
Writ. Abil.					1.0	.85	.67	.52	66	.52	.54
TOEFL Total						1.0	.79	.69	.78	69	.69
Accept. Cloze							1.0	.75	.94	.76	.72
Dict. Total								1.0	.69	96	.93
Exact Cloze									1.0	.70	69
Easy Dict.										1.0	.85
Hard Dict.											1.0

(entnommen aus: Irvine/Atai/Oller, 1974:249)

Auch hier wird wieder geschlossen, daß der CT höhere Korrelationen mit integrativen Testverfahren aufweist.

> "The patterns of intercorrelation do not support the attempt to separate skills and components of skills on the TOEFL. They argue rather in favour of language tests that are task oriented once that require the pragmatic use of language for communicative purposes." (Irvine/Atai/Oller, 1974:251)

Die Ergebnisse beider Clozeauswertungsverfahren wurden miteinander korreliert und ergaben einen Koeffizienten von 0,94. Die Ähnlichkeit der Ergebnisse bei den beiden Auswertungsverfahren läßt es geraten erscheinen - so wurde im Einklang mit Stubbs und Tucker (siehe unten) gefolgert - bei non-native speakers die 'exact scoring'- Methode zu verwenden, um Schwierigkeiten bei der Bestimmung der Akzeptabilität zu vermeiden (cf. 2.4.3.) Dabei werden aber die pädagogischen Implikationen dieses Verfahrens übersehen

Auch Stubbs und Tucker (1974) untersuchten das Verhältnis zwischen der 'exact scoring'- Methode und der 'acceptable word'- Auswertungsmethode und stellten eine Korrelation von 0,97 zwischen beiden fest, weshalb erstere als die einfachere empfohlen wurde. Der CT aus insgesamt 50 Lücken im Abstand von 5 Worten wurde bei insgesamt 211 Studenten an der Universität von Beirut, von denen 155 Arabisch und der Rest andere Muttersprachen hatten, durchgeführt. Als Validitätskriterium diente das 'English

Entrance Examination' (EEE) der Amerikanischen Universität in Beirut. Dieses EEE bestand aus vier Untertests: Grammatik, Vokabular und zwei nicht näher beschriebene Lesetests.

Da Korrelationen zwischen dem CT und Gesamt-EEE von 0,71 ('exact scoring') und 0,76 ('Acceptable scoring') errechnet wurden, schließen die Autoren, daß der CT ein valider Test der englischen Fremdsprachenkompetenz ist. Obwohl die 'acceptable word'- Werte konsistent größere Korrelationskoeffizienten bei den Untertests des EEE erbrachten (bei Grammatik 0,70 bzw. 0,66; bei Vokabular 0,65 bzw. 0,60; bei Lesen 0,70 bzw. 0,67), wurde die 'exact scoring'- Methode wegen ihrer Einfachheit für non-native Lehrer empfohlen.

Wegen ihres Bezugs zum Elementarbereich sei eine Studie von Wijnstra & Wageningen (1974) erwähnt. Es handelt sich allerdings um spanische Kinder, die in den Niederlanden Holländisch lernten. 102 Schüler der Grundschulklassen 4 - 6 bearbeiteten 45 Aufgaben eines CT. Dabei wurden zwei verschiedene Auswertungsmodalitäten verglichen: Ein multiple-choice-Verfahren und das 'acceptable word'- Verfahren. Es wurden mit der KUDER-RICHARDSON-FORMEL 20 Reliabilitätskoeffizienten von 0,88 für das multiple-choice-Verfahren und 0,89 für das 'acceptable word'- Verfahren errechnet.

Die Studie von Pike (1973) untersuchte

> "how well several component language competencies are estimated by the TOEFL subtests and by alternative experimentalmeasures." Pike (1973:

Zu den 'alternative experimental measures' gehörten ein Standard CT und ein Clozentropyverfahren (nach Darnell, 1968) mit 25 Aufgaben. Die insgesamt 442 Probanden setzten sich zusammen aus 98 peruanischen, 145 chilenischen und 199 japanischen Universitätsstudenten, die in Lima bzw. Santiago und Tokio getestet wurden. Die mit der KUDER-RICHARDSON-20-FORMEL berechneten Reliabilitätskoeffizienten lagen zwischen 0,78 - 0,91 für 'exact scoring' und zwischen 0,82 und 0,85 für das Clozentropyverfahren. Von besonderem Interesse sind jedoch die zur Validierung vorgenommenen Korrelationsberechnungen zwischen den CT und den Clozentropyverfahren einerseits und den TOEFL Subtests andererseits:

TAFEL 8

entnommen aus: Pike, 1973:99

	PROBANDEN AUS	HÖRVERST.	GRAMM.	WORTSCH.	LESEVERST.	SCHREIB-FÄH.
CLOZENTROPY	Peru	81	83	82	79	82
	Chile	68	82	70	76	76
	Japan	52	64	64	69	64
CLOZE	Peru	80	78	83	79	80
	Chile	65	79	70	77	76
	Japan	51	62	64	65	63

Neben der Tatsache, daß hier kein klares Verteilungsmuster der höheren Korrelationen bei integrativen Subtests und der weniger hohen bei 'discrete-point'- Subtests besteht, ist vor allem die Tatsache zu unterstreichen, daß Probanden mit verschiedenen Nationalitäten und Muttersprachen und unterschiedlichen soziokulturellem Hintergrund verschieden eingestuft werden.
Dies unterstützt die eingangs getroffene Feststellung, daß Forschungsergebnisse nicht einfach auf andere Verhältnisse übertragen werden können.

Eine ähnliche Beobachtung wurde von Hisama (1976) gemacht. Der verwendete CT entsprach allerdings nicht ganz dem hier zu untersuchenden Standardmodell, da nicht jedes n-te (7te) Wort ausgelassen wurde, sondern wie Hisama beschreibt:

> "The approach followed in this study can be called the variable-ratio approach where the intervals vary around n. This change allowed us greater freedom in designing the New Cloze Test. For example, the test was designed so that the total number of words to be deleted was well balanced in terms of function words and content words." Hisama (1976:96)

Wenngleich wegen dieser Änderung nicht näher auf die Gesamtstudie eingegangen werden soll, ist die Verteilung der Ergebnisse beim CT und den anderen Vergleichstests (CELT-S = Comprehensive English Language Test for Speakers of English as a Second Language - Structure; CELT-L = Comprehensive English Language Test for Speakers of English as a Second Language - Listening; RFU = Reading for Understanding Test) von besonderem Interesse. Dies ist ersichtlich aus der folgenden Darstellung der Leistungsprofile:

ABB 4

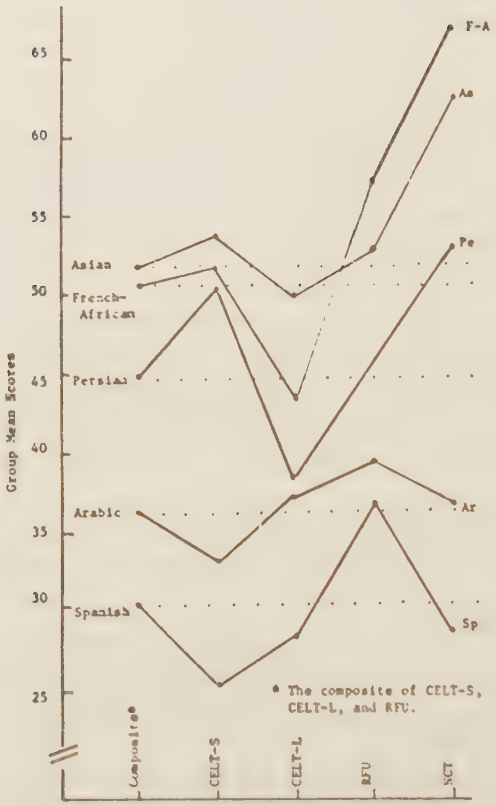


FIGURE 4-1 PROFILES OF DIFFERENT LANGUAGE GROUPS

(entnommen aus: Hisama, 1976:116)

Die oben gemachte Beobachtung der verschiedenen Reaktionen auf Tests durch Probanden verschiedener sprachlicher und soziokultureller Hintergründe wird hierdurch wieder bestätigt.

Weiterhin ist es auffällig, daß zwei Grundmuster der Profile vorhanden sind. Die 'Asian-French-African' und 'Persian'- Probanden zeigen ebenso wie die 'Arabic' und 'Spanish'- Probanden relativ ähnliche Muster. Nach dieser Darstellung verhalten sich Probanden oberen und unteren Kompetenzniveaus jeweils ähnlich. Auch das Kompetenzniveau kann also einen Einfluß auf das Leistungsprofil haben.

Kržyzanowsky (1976) führte eine Batterie von Englischtests bei 94 polnischen Sekundarschülern durch mit dem Ziel, den CT als Meßinstrument der englischen Fremdsprachenkompetenz zu erproben. Die Probanden wurden in drei Gruppen eingeteilt. Die Englischkenntnisse der Gruppen I (N=31) und II (N=31) wurden als intermediate, die der Gruppe III (N=32) als advanced bezeichnet. Die Testbatterie bestand aus einem CT, bei dem jedes 7. Wort ausgelassen wurde, mit insgesamt 50 Lücken, einem Grammatiktest (structure test), einem Wortschatztest und einem Diktat. Bei der Bewertung wurde jedes akzeptable Wort gezählt.
Folgende Reliabilitätsfaktoren und Testwerte wurden errechnet:

TAFEL 9

	GRUPPE	RANGE	ARITHM. MITTEL X	STANDARD-ABWEICHUNG	RELIABILITÄT	MITTLERER SCHWIERIGKEITSGRAD (FACILITY)
N=31	I	36	22,1	9,5	0,86	0,44
N=31	II	33	24,4,	9,7	0,87	0,49
N=32	III	22	39,1	5,5	0,72	0,78
N=94	I II III	44	28,4	11,4	0,905	0,57

Für den CT wurden folgende Korrelationen mit dem Ganzen und den Subtests errechnet:

TAFEL 10

Group	Reliability	CLOZE I	VOCAB	STRUCT	DICTAT	TOTAL
I	.86		.89	.88	.86	.945
II	.87	CLOZE I	.90	.79	.75	.95
III	.72		.65	.49	.78	.89
P	.905		.89	.88	.88	.95
I	.92			.85	.89	.98
II	.90		VOCAB	.91	.79	.98
III	.79			.76	.76	.875
P	.95			.85	.90	.97
I	.85				.90	.95
II	.79			STRUCT	.71	.95
III	.85				.78	.91
P	.89				.855	.93
I	.87					.96
II	.88				DICTAT	.88
III	.86					.935
P	.92					.96
I	.96					
II	.95					TOTAL
III	.92					
P	.975					

(entnommen aus: Kržyzanowski, 1976:34)

Dabei wiesen die Gruppen I, II und III eine gemeinsame Varianz von 90% auf. Diese Resultate sind beachtenswert, aber es wäre interessant zu sehen, wie die Korrelationen beeinflußt werden, wenn der CT aus der Testbatterie genommen wird, bevor die Analyse durchgeführt wird. Was die Subtests betrifft, werden frühere Ergebnisse höherer Korrelationen bei sog. integrativen Tests, wie z.B. Diktat, nicht bestätigt.

Kržyzanowsky erklärte diese Tatsache damit:

> "The fact that both Dictation for group III and Vocabulary for the rest of the test subjects had the highest reliability out of the three tests correlated seems to be the most important [contributing factor] here."
> Kržyzanowsky (1976:35)

Mit anderen Worten: der CT korrelierte mit dem Subtest, der den höchsten Reliabilitätsfaktor aufweist, für jede Gruppe von Probanden am besten.

Akzeptiert man diese Erklärung, dann muß man schließen, daß die Reliabilität des Kriteriums die Korrelationskoeffizienten beeinflußt, was nicht ohne Folgen bleiben darf bei der Auswahl geeigneter Validitätskoeffizienten.

Kržyzanowsky erweitert diese Studie durch die Einbeziehung des Lehrerurteils als Validitätskriterium. Die Lehrer beurteilten die Schüler nach ihrer gesamten Fremdsprachenkompetenz. Die Ergebnisse wurden mit den Testergebnissen korreliert und in folgender Tabelle niedergelegt:

TAFEL 11

Correlation Between Test Results and the Teachers' Classification

Tests	Group I	Group II	Group III
CLOZE I	.87	.88	.77
VOCABULARY	.895	.91	.76
STRUCTURE	.83	.88	.86
DICTATION	.87	.76	.73
TOTAL	.90	.915	.89

(entnommen aus: Kržyzanowsky, 1977:36)

Überraschend sind die hohen Werte im Bereich des Vokabeltests der Gruppe I und II und des Grammatiktests der Gruppe III. Wegen fehlender genauerer Angaben sind weitere Interpretationen nur schwer möglich. Allgemein schließt Kržyzanowsky jedoch:

> "These data are very important for the estimation of the general validity of the whole battery of tests. However, we should remember that, though aiming at the same objective, the two sources of data (i.e. 'teachers' lists and test statistics) are of apparently different character and, as such are not fully comparable. On the one hand, the teachers' classifications represent a much deeper and more integrated knowledge of learners' proficiency than that offered by the tests, but on the other hand they are not formalized, being thus less precise and less reliable."
>
> Kržyzanowsky (1977:36)

Dieser Aussage stehen einige Argumente entgegen:

Auch wenn die Beurteilungen (Lehrerurteil und Testergebnis) auf verschiedenem Wege entstehen, können beide numerisch quantifizierbar und vergleichbar gemacht werden. Es ist bedauerlich, daß dieses hier nicht gemacht worden ist. Was die Qualität des Lehrerurteils betrifft, so sagt Ingram dazu, daß die Beurteilung von Schülern durch erfahrene Lehrer

> "appears to be one of the best criterion measures of immediate validity as long as it is limited to one class and one teacher at a time. When a series of small scale comparisons between test scores and teachers' ratings show sufficient agreement, there is good reason to be satisfied at least about immediate validity."
> Ingram (1977:21)

Als allgemeines Ergebnis der Studie stellt der Autor die sehr gute Eignung des CT-Verfahrens zur Messung einer allgemeinen Fremdsprachenkompetenz fest, insbesondere als Teil einer Testbatterie.

1976 führte Hinofotis (1976) eine Untersuchung an insgesamt 107 Studenten des Southern Illinois University Center for English as a Second Language durch. Die Studenten umfaßten ein weites Spektrum von sprachlichen und soziokulturellen Hintergründen bei einer großen Bandbreite der fremdsprachlichen Kompetenz. Der durchgeführte CT bestand aus insgesamt 50 Lücken im Abstand von 7 Wörtern. Die Auswertung erfolgte nach der 'exact scoring' und der 'acceptable word'- Methode. Folgende Werte wurden gemessen:

TEST	AUSWERTUNGSMETHODE	MÖGL. PUNKTE	ARITHM. MITTEL	STANDARD ABWEICH.	N ZAHL D. PROBAND.	RELIABILITÄTSKOEFF.	BERECHNET NACH KUDER-RICH.-FORMEL
CLOZE	EX	50	11,9	2,08	107	0,61	
	AC	50	15,3	7,30	107	0,85	

TAFEL 12

(entnommen aus: Hinofotis, 1976:94)

Zur Validierung wurden die CT-Ergebnisse mit einer Reihe etablierter Testverfahren korreliert und folgende Werte erzielt:

TAFEL 13

	Cloze exact	Cloze acceptable	N
TOEFL	6,71	0,79	52
Center of English as a Second Lang. Placement Test	0,80	0,84	107
Foreign Service Interview mit versch. Interviewern	0,70 (Redd)	0,72	59
	0,48 (Perkins)	0,54	47
	0,60 (All)	0,63	106

(entnommen aus: Hinofotis, 1976:106)

Die gefundenen Korrelationen mit den Subtests dargestellt in der folgenden Tabelle unterstützen die These, daß der CT mit integrativen Tests eher korreliert als mit discrete-point Tests.

TAFEL 14

	Cloze Exact	Cloze Acceptable	Redden Cloex	Redden Cloac	Perkins Cloex	Perkins Cloac
TOEFL (N=52)						
Listening Comp	.47	.51				
Structure	.51	.58				
Vocabulary	.59	.62				
Reading	.68	.77				
Writing	.55	.64				
CESL Placement (N=107)						
Listening Comp	.71	.73				
Structure	.63	.69				
Reading	.80	.80				
FSI (N=106)			(N=59)		(N=47)	
Accent	.50	.51	.51	.53	.45	.47
Grammar	.60	.62	.72	.73	.45	.52
Vocabulary	.53	.55	.58	.59	.46	.51
Fluency	.56	.59	.70	.71	.40	.48
Comprehension	.54	.58	.59	.62	.47	.53

Pearson Product Moment Correlations for Cloze Tests (Both Scoring Methods) with Subtest Scores of the TOEFL, CESL Placement, and FSI Examinations. Correlations for all Subjects and by Subfiles are Included for the FSI Subtests. $p < .005$ for all Correlations.

(entnommen aus: Hinofotis, 1976:104)

Insgesamt wird geschlossen, der Cloze Test

"appears to be a valid measure of overall ESL /English as a Second Language/ proficiency."
Hinofotis (1976:107)

Alderson (1978) berichtet von einer äußerst vielschichtigen Studie, die den CT bei native speakers und non-native speakers in Abhängigkeit von einer großen Anzahl von Variablen untersucht (siehe auch Alderson, Juni 1980). Dazu gehören 12 verschiedene CTs gebildet durch vier verschiedene Lückenfrequenzen (6, 8, 10 und 12) und drei verschiedene Textschwierigkeiten, fünf verschiedene Auswertungsmethoden,

1) Exact Scoring (Exact)
2) any semantically acceptable word (Semac)
3) any grammatically correct word (Grco)
4) any word from the same form class as the deleted word (Idfc)
5) any word fullfilling the same grammatical function as the deleted word. (Acfc)

Insgesamt nahmen 360 erwachsene Englischlerner an der Untersuchung teil. Aufgeteilt in Gruppen von 30 Studenten bearbeiteten sie jeweils eine Testkombinationsmöglichkeit, die durch ein kompliziertes Rotationsverfahren verteilt wurden.
Das Ziel dieser Studie liegt vor allem in der Beantwortung methodischer Detailfragen. Wegen der Vielfalt der ineinander verzahnten und voneinander abhängigen Variablen werden allgemeine Feststellungen sehr erschwert.
Was die Reliabilität des Tests anbetrifft wird ausgesagt, daß eine Berechnung mit der KR-20-FORMEL für die grammatisch und semantisch akzeptable Auswertungsmethode Werte zwischen 0,74 und 0,92 berechnet wurden. Die 'exact scoring'- Methode bringt in allen Fällen niedrigere Koeffizienten als die anderen Methoden. Die Textschwierigkeit läßt keine konsistenten Vorteile größerer oder geringerer Schwierigkeitsgrade erkennen.[20b] Es ist nicht der Fall, daß bei einem Zuwachs des Kontextumfanges, das heißt bei größerer Lückenfrequenz, die Resultate reliabler werden. Ein

20b Vgl. Alderson, 1980:72

Wechsel von Lückenfrequenz läßt kein beobachtbares Muster der Reliabilität erkennen (1978:353). Als Validitätskriterium diente die English Language Battery (ELBA) bestehend aus zwei Teilen, 'Listening' und 'Reading', im Multiple-choice-Format. Teil eins (Pt I = Listening) hat vier Untertests:

1) Sound Recognition
2) Intonation
3) Stress
4) Listening Comprehension.

Teil zwei (Pt II) hat drei Untertests:

5) Grammatik
6) Vocabulary
7) Reading Comprehension.

Die Korrelationen zwischen den CT und seinen verschiedenen Auswertungsmethoden und dem ELBA und seinen Untertests sind in TAFEL 15 niedergelegt. Herausgegriffen sei hier insbesondere, daß die 'semantic acceptable word'-Bewertungsmethode (Semac) konsistent bessere Werte zeigt. Für den Gesamt-ELBA-Test bewegen sie sich zwischen 0,71 und 0,80 gegenüber 0,61 und 0,67 bei der 'exact scoring'- Methode.

Was die Korrelationen mit den Subtests betrifft, die ebenfalls aus Tafel 15 zu entnehmen sind, werden in dieser Studie die Ergebnisse von Darnell (1968) bestätigt, daß Cloze am niedrigsten mit 'Reading Comprehension' und am höchsten mit 'Vocabulary' und 'Grammar' korreliert. Dies steht im Widerspruch zu anderen Ergebnissen, z.B. Oller/Conrad (1971) und Hinofotis (1976).

Alderson führt auch eine Faktorenanalyse durch mit dem Ziel, die Frage zu beantworten, ob Cloze eine allgemeine Fremdsprachenkompetenz bzw. einen Faktor dieser mißt, und gibt folgende Antwort:

> "In answer to the initial question as to whether cloze measures the same 'proficiency in English as a foreign language' factor as the ELBA, the evidence from this study is that, on the whole, the cloze is such a measure. The unifactorial solution shows cloze to load high on a general proficiency factor. When two factors are extracted, the cloze is associated with the core proficiency, and even when five factors are extracted, the cloze is still seen to associate with this factor. In addition, the study confirms some studies and refutes others to show that the any-acceptable-word scoring procedure is the best measure of the proficiency factor, but, further that almost any scoring procedure is capable of measuring this factor." Alderson (1978:341)

Alderson differenziert später dieses Ergebnis noch weiter in Bezug auf 'high and low order skills' (vgl. Alderson, 1980:105).

TAFEL 15

(entnommen aus: Alderson, 1978:278)

Pearson Product Moment correlations of cloze scores with ELBA and dictation by test.

a) Difficult Text

	ELBA										Dictation	
	1	2	3	4	Pt1	5	6	7	Pt2	Total	I	II
Cloze												
Exact	.36	.49	.25	.55	.51	.71	.57	.54	.69	.66	.50	.54
SEMAC	.50	.57	.30	.67	.65	.79	.73	.63	.82	.80	.58	.59
CRCO	.37	.49	.35	.50	.51	.72	.61	.50	.71	.66	.52	.51
IDPC	.41	.53	.34	.57	.56	.74	.64	.50	.73	.71	.52	.52
ACPC	.39	.50	.33	.56	.55	.71	.64	.50	.72	.69	.51	.51

(ELBA: Cloze, n = 81; Dict I: Cloze, n = 87; Dict II: Cloze, n = 88. $p < .001$)

b) Medium Text

											Dictation	
	1	2	3	4	Pt1	5	6	7	Pt2	Total	I	II
Cloze												
Exact	.49	.36	.36	.56	.63	.58	.55	.51	.64	.67	.58	.58
SEMAC	.57	.47	.43	.70	.76	.65	.63	.59	.72	.78	.69	.69
CRCO	.59	.36	.37	.66	.73	.62	.61	.55	.69	.75	.62	.66
IDPC	.42	.33	.44	.52	.57	.54	.49	.49	.58	.62	.51	.56
ACPC	.43	.29	.38	.52	.57	.53	.48	.50	.58	.61	.51	.56

(ELBA: Cloze, n = 91; Dictation: Cloze, n = 95 $p < .001$)

c) Easy Text

											Dictation	
	1	2	3	4	Pt1	5	6	7	Pt2	Total	I	II
Cloze												
Exact	.47	.47	.29	.63	.58	.58	.55	.47	.59	.61	.70	.58
SEMAC	.55	.53	.45	.70	.68	.69	.61	.58	.69	.71	.76	.67
CRCO	.51	.44	.42	.65	.62	.66	.59	.57	.67	.67	.74	.66
IDPC	.43	.43	.35	.57	.54	.54	.46	.47	.54	.56	.64	.58
ACPC	.43	.39	.35	.58	.54	.56	.48	.49	.56	.57	.69	.62

(ELBA: Cloze, n = 92 Dictations: Cloze, n = 93 $p < .001$)

Die jüngste Studie stammt von Wainman (1979), der den CT mit 205 'final-year secondary' und 'first-year university undergraduate students' untersuchte. Er verglich zwei verschiedene CTs mit jeweils 40 Aufgaben, die einmal in einem 10-Wort-Lückenabstand gesetzt wurden und zum anderen nach der Intuition des Entwerfenden nur aus 'content-words' bestanden. Ergebnis der Untersuchung war die Feststellung, daß die Resultate ('mean scores') in beiden Verfahren keine wesentlich unterschiedlichen Informationen liefern. Als Kriterium zur konkurrenten Validierung wurden Lehrerbeurteilungen in 'Reading Comprehension' und 'Writing' mit dem CT korreliert. Die Korrelationskoeffizienten lagen für 'Reading Comprehension' zwischen 0,65 und 0,74 und für 'Writing' zwischen 0,41 und 0,68.

Um eine Übersicht und bessere Teilvergleiche zu ermöglichen, sind die Ergebnisse der angeführten Studien in der Tabelle TAFEL 16 zusammengefaßt:

TAFEL 16a) STUDIEN ZUM STANDARDMODELL DES CLOZE TESTS ALS MESSINSTRUMENT DER FREMDSPRACHENKOMPETENZ :

R E L I A B I L I T Ä T S K O E F F I Z I E N T E N

Studie	N = Größe d. Stichprobe	Art der Probandengruppe	n = Anzahl d. Testaufgaben	Berechnung der Reliabilität durch	Auswertungsmethode	Reliabilitäts-koeffizient
Darnell (1970)	48	Universität (ESL)	50	Hoyt	CLZNT	0,86
Anderson (1970)	60	Elementary pupils in Papua, New Guinea (ESL)	50	Split half	AC	0,95
Oller (1972)	398	University (ESL)	50	K-R 20	EX AC	0,80-0,92 0,90-0,95
Pike (1973)	442	University Peru, Chile, Japan	25	K-R 20	EX CLZNT	0,78-0,91 0,82-0,85
Wijnstra & Wageningen (1974)	102	Schülerjahrg. 4 - 6, Holland	45	K-R 20	MC AC	0,88 0,89
Krzyzanowsky (1976)	94	Sekundarschule	50	K-R	AC	0,86-0,90
Hinofotis (1976)	107	University (ESL)	50	K-R 20	EX AC	0,61 0,85
Alderson (1978)	360	University (ESL)	ca 50	K-R- 20	EX AC	niedriger als bei AC 0,74-0,92

TAFEL 16b) STUDIEN ZUM STANDARDMODELL DES CLOZE TESTS ALS MESSINSTRUMENT DER FREMDSPRACHENKOMPETENZ:

VALIDITÄTSKRITERIEN

Studie	N = Größe d. Stichprobe	Art der Probandengruppe	Externes Validitätskriterium	Auswertungsmethode	Validitätskoeffizient	Determinationskoeffizient
Darnell (1970)	48	Universität (ESL)	TOEFL	CLZNT	0,84	0,71
Anderson (1970)	[illegible]		Watt's Reading Test (selbstentw.)	AC	0,78-0,85	0,61-0,72
Oller & Conrad (1971)	102	[illegible] (ESL)	ESLPE, UCLA	EX	0,88	0,77
Oller (1972)	398	[illegible] (ESL)	ESLPE, UCLA	EX AC	0,73-0,87 0,80-0,89	0,53-0,76 0,64-0,79
Irvine&Atai& Oller (1974)	159	Universität	TOEFL	EX AC	0,78 0,79	0,61 0,62
Stubbs & Tucker (1974)	211	[illegible] (ESL)	English Entrance Examination, American Univ., Beirut	EX AC	0,71 0,76	0,50 0,58
Krzyzanowsky (1976)	94	Sekundarschule Polen	selbstentw. Testbatterie incl. CT Lehrerurteil	AC	0,95 0,89-0,90	0,90 0,79-0,81

STUDIEN ZUM STANDARDMODELL DES CLOZE TESTS ALS MESSINSTRUMENT DER FREMDSPRACHENKOMPETENZ:

V A L I D I T Ä T S K R I T E R I E N

(Fortsetzung der TAFEL 16 b)

Studie	N = Größe d. Stichprobe	Art der Probandengruppe	Externes Validitätskriterium	Auswertungsmethode	Validitätskoeffizient	Determinationskoeffizient
Hinofotis (1976)	52	Universität (ESL)	TOEFL	EX	0,71	0,50
				AC	0,79	0,62
	107	Universität (ESL)	Placement and Test Center for ESL (CESL) Southern Illinois University	EX	0,80	0,64
				AC	0,84	0,71
	106	Universität (ESL)		EX	0,60	0,36
				AC	0,63	0,40
Alderson (1978)	360	Universität (ESL)	English Language Battery	EX	0,61-0,67	0,37-0,45
				AC	0,71-0,80	0,50-0,64
Wainman (1979)	205	Sekundarschule/ Universität	Lehrerbeurteilung 'Reading Comprehension' 'Writing'	AC	0,65-0,74 0,41-0,68	0,42-0,55 0,17-0,46

E r k l ä r u n g / L e g e n d e :

KR-20	KUDER-RICHARDSON-FORMEL 20
CLZNT	Clozentropy Methode
AC	'any acceptable word'- Methode
EX	'exact scoring'- Methode
MC	Multiple Choice
ESL	English as a Second Language

In der Gesamtübersicht lassen sich folgende allgemeinen Feststellungen machen und Probleme erkennen:

1) Zur Reliabilität:

Die vorliegenden Untersuchungen weisen Reliabilitätskoeffizienten zwischen 0,61 und 0,95 auf. In allen Fällen werden diese Ergebnisse als akzeptable Werte angesehen und der CT als reliables Meßinstrument erkannt. Bis auf zwei Studien wurde die Reliabilität immer mit der KUDER-RICHARDSON-20-FORMEL berechnet. Die Formel von Hoyt (siehe Lienert, 1967: 228) ist jedoch nur eine Umgestaltung der KUDER-RICHARDSON-20-FORMEL. Hier müßte eine gewisse Skepsis angemeldet werden. Nach Lienert (1967: 227) bedingt diese Formel die Annahme, daß alle Aufgabeninterkorrelationen etwa gleich sind. Keine der Untersuchungen macht jedoch direkte Aussagen über diese Frage der Aufgabeninterkorrelation. Indirekt darauf bezieht sich die Feststellung von Alderson, daß Cloze "sentence bound" ist, in anderen Worten, nur vom unmittelbaren Kontext des Satzes abhängig ist. Kann dies angenommen werden, dann sind die Einzellücken weitgehend voneinander unabhängig und die Verwendung von KR-20 wäre gerechtfertigt. Da berechtigter Zweifel besteht über die gegenwärtige Unabhängigkeit der Aufgaben, bedarf diese Frage noch der Klärung (siehe z.B. Chihara et al, 1977)

Die Art der Probanden in Bezug auf ihr fremdsprachliches Kompetenzniveau, ihren muttersprachlichen und soziokulturellen Hintergrund beeinflußt die Reliabilität, wodurch Forschungsergebnisse schwer vergleichbar und nicht übertragbar auf andere Probandengruppen werden.

Verschiedene andere Variablen des Tests beeinflussen die Reliabilität:

Die Testauswertungsmethode, die Anzahl der Testitems und die Testschwierigkeit müssen bei der Testkonstruktion berücksichtigt werden und sollen deshalb noch gesondert untersucht werden (siehe Kapitel zu Methodische Aspekte des CT).

2) Zur Validität:

Tabelle 2 weist Validitätskoeffizienten zwischen 0,41 und 0,90 auf. Dabei bewegen sich die meisten im Bereich zwischen 0,70 und 0,80, was Determinationskoeffizienten von 0,49 bis 0,64 entspricht. Alle Autoren folgern daraus, daß die gemeinsame Varianz zwischen den Vergleichskri-

terien groß genug ist, um auf Übereinstimmung der Testergebnisse der beiden Meßinstrumente schließen zu können. In den meisten Fällen sind die Vergleichskriterien standardisierte, etablierte Fremdsprachentests. In einigen Fällen werden selbstentworfene Tests oder Lehrerbeurteilungen als Vergleichskriterien verwendet, von denen man sicher eine gewisse Gültigkeit erwarten kann.
Insgesamt wird deshalb auf eine ausreichende Gültigkeit des CT als Meßinstrument einer allgemeinen Fremdsprachenkompetenz geschlossen.
Wirft man jedoch einen Blick auf die Korrelationen mit den Subtests, werden widersprüchliche Aussagen gemacht. So schließen z.B. Darnell (1968) und Alderson (1978) aus den Subtestkorrelationen, daß Cloze am niedrigsten mit 'Reading Comprehension Tests' und am höchsten mit 'Vocabulary' und 'Grammar Tests' korreliert. Oller /Conrad (1971) und Hinofotis (1976) z.B. jedoch widersprechen diesen Aussagen völlig mit der Feststellung, daß Cloze am meisten gerade mit integrativen Tests, wie z.B. 'Reading Comprehension' korreliert. Eine mögliche Erklärung könnte in der Verschiedenartigkeit von jeweiligen Grammatik- bzw. Reading Comprehension Tests der einzelnen Testbatterien liegen.
Eine andeutungsweise Erklärung gibt auch Krzyzanowsky (siehe oben), der unterschiedliche Reliabilitätskoeffizienten der Subtests verantwortlich macht für Validitätsschwankungen. Dies ist nicht unlogisch, wenn man in Betracht zieht, daß die Reliabilität eine Vorbedingung der Validität ist. Aus den Vorlagen sind jedoch kaum die Reliabilitätskoeffizienten der Validierungskriterien und der Subtests ersichtlich. Obgleich standardisierte Tests wegen der von ihnen beanspruchten hohen Gültigkeit als Validitätskriterien verwendet werden, ist dies nicht unproblematisch.

Nicht nur die Reliabilität sondern auch die Validität wird durch die A r t der P r o b a n d e n g r u p p e beeinflußt. Die oben erwähnten Studien von Pike (1973) und Hisama (1976) zeigen deutlich unterschiedliche Ergebnisse im kontrastiven Vergleich verschiedener Sprachgruppen bzw. soziokultureller Hintergründe.

Auch das L e i s t u n g s n i v e a u ist ein Faktor, der die Validität beeinflußt. So zeigen verschiedene Niveaugruppen in der Studie von Hisama (1976) verschiedene Leistungsprofile, wodurch Unterschiede der Korrelationskoeffizienten verursacht werden.

2.4.3. Forschung zu methodischen Apekten des Cloze Test

Wie oben dargestellt werden Reliabilität und Validität von verschiedenen methodischen Variablen beeinflußt wie

Auswertungsverfahren
Lückenfrequnz
Anzahl der Aufgaben
Textschwierigkeit und
Art der Aufgaben.

Diese Variablen beeinflussen sich gegenseitig (so ist es ohne weiteres einsichtig, daß z.B. ein kürzerer Lückenabstand im gleichen Text gleichzeitig eine größere Aufgabenanzahl, damit eine breitere Streuung, bringt, was für den Probanden schwieriger zu lösen ist); sie sind auch voneinander abhängig.

Trotzdem sollen hier soweit möglich die einzelnen Variablen isoliert betrachtet werden.

Die Auswertungsmethode ('Scoring')

Seit Taylor (1953) die 'exact scoring method', bei der nur das mit dem Originaltext übereinstimmende Wort gewertet wird, für native speakers empfohlen hat, sind mehrere andere Methoden entwickelt und erprobt worden. Besonders im fremdsprachlichen Bereich erschien die 'exact scoring'- Methode nicht immer zufriedenstellend.
Vom pädagogischen Aspekt gesehen erscheint es nachteilig, jene Lösungen, die vom Kontext her akzeptabel sind, nicht anzuerkennen. Die Probanden fühlen sich ungerecht behandelt und sind frustriert, wenn ihre Anwort nicht direkt in ihre Punktezahl einfließt, nur weil sie nicht genau dem Originaltext entspricht (vgl. auch 4., S. 124).
Eine Methode, die öfter erprobt worden ist, ist die oben erwähnte, von Darnell (1968) entwickelte 'Clozentropy'- Methode. Sie basiert - vereinfacht ausgedrückt - auf Logarithmen der Häufigkeit, mit der Vergleichsgruppen von native speakers die Clozeaufgaben beantworten (genauer vgl. Reilly, 1971:350-353, cf. 2.4.3.). Mit ihr wurden günstige Reliabilitäts- und Validitätskoeffizienten erzielt (siehe TAFELN 16a und 16b zu den Studien von Darnell und Pike). Übereinstimmend sahen die Autoren einen großen Nachteil in der Praktikabilität der Methode.
Oller (1972) untersuchte fünf Auswertungsmethoden mit 398 ESL-Studenten. Seine Methode bestand in unterschiedlicher Gewichtung für verschiedene Antworten:

M_1: nur das gleiche Wort wie im Originaltext wird gezählt;
M_2: gleiche Wörter und kontextuell akzeptable Antworten werden gez.;
M_3, M_4 und M_5 gaben verschiedene Gewichtungen für gleiche Wörter, akzeptable Antworten, Antworten, die im Hinblick auf

den weiteren Kontext akzeptabel sind ('long-range-acceptability') und Antworten, die nur im unmittelbaren Kontext akzeptabel sind 'short-range-acceptability');

M_3: 4 3 2 1;
M_4: 2 2 2 1;
M_5: 2 2 1 1.

Da er keine signifikanten Unterschiede zwischen M_3, M_4 und M_5 fand, was im Hinblick auf das Gewichtungssystem nicht überraschend ist, und da die Methoden sehr komplex sind, lehnte er sie ab und konzentrierte sich auf M_1 und M_2. Er fand, daß Validitätskoeffizienten bei der 'acceptable word'-Methode konsistent höher waren als bei der 'exact word'- Methode (siehe Tabellen 16a und 16b).

Das gleiche gilt für die Reliabilitätskoeffizienten. Oller schloß daraus, daß für non-native speakers die 'acceptable word'- Methode vorzuziehen ist.

Oller, Atai und Irvine (1974) untersuchten wiederum die 'exact scoring'-Methode sowie die Methode des 'acceptable scoring' bei ESL-Proficiency Tests und stellten fest, daß beide Methoden mit 0,94 korrelierten. Darüberhinaus waren die Korrelationen mit den externen Validitätskriterien fast gleich. Im Gegensatz zu Oller (1972) wurde geschlossen, daß die 'exact-scoring'-Methode wegen ihrer einfacheren Durchführung vorzuziehen ist. Dieses Ergebnis wurde von Stubbs und Tucker (1974) bestätigt, die ebenfalls den Gebrauch der 'exact scoring'- Methode empfahlen, weil sie mit 0,97 mit der 'acceptable word'- Methode korrelierte. Sie vernachlässigten dabei jedoch die Tatsache, daß die 'accepable word' Methode konsistent höhere Validitätskoeffizienten liefert (siehe Tabelle 16 b).

Bessere Ergebnisse sowohl der Reliabilität als auch der Validität liefert auch die Untersuchung von Hinofotis (1976) für die 'acceptable word'-Methode gegenüber der 'exact scoring'- Methode.

In einer Studie, die besonders methodische Probleme untersucht, vergleicht Alderson (1978) fünf verschiedene Methoden. Neben den erwähnten Methoden untersucht er sog. 'grammatical scoring procedures', um die Sensibilität des CT im Bereich der Syntax zu messen.

1) Same form class: diese Methode vernachlässigt Kontextzusammenhänge und akzeptiert alle Lösuungen, die der syntaktischen Formklasse des Originals entsprechen. Der Begriff der Formklasse wurde dabei nach Fries (1952) definiert. Alderson erkennt selbst Schwierigkeiten dieser Methode.

Insbesondere können Lücken in durchaus akzeptabler Weise mit Wörtern verschiedener Formklassen gefüllt werden.

2) Acceptable Form Class, same function: diese Methode vermeidet die aufgeführte Schwierigkeit. Der Grammatikalität der Antwort wird jedoch auch durch diese Methode nicht voll Rechnung getragen.

3) Grammatically correct response: Diese Methode berücksichtigt auch das letztere Problem.

Die Ergebnisse des statistischen Vergleichs aller 5 Methoden faßt Alderson so zusammen:

> "There is comparatively high agreement among scoring procedures as to ranking of non-native speaker subjects; in other words, they tend to measure the same thing, or at least in a similar manner." Alderson (1978:240)

Darüber hinaus ergibt die 'acceptable word'- Methode auch bei Alderson die höheren Reliabilitäts- und Validitätskoeffizienten. Er schließt, daß die 'acceptable word'- Methode vorzuziehen sei. Dasselbe Ergebnis erbringt auch noch eine Studie von Porter (1978). N = 39 Studenten der Universität Krakau. Die Reliabilitätskoeffizienten bewegen sich zwischen 0,39 und 0,57 für 'exact scoring' und zwischen 0,57 und 0,65 bei der 'acceptable word'- Methode.

Trotz offensichtlicher statistischer Vorteile der 'acceptable word'- Methode besteht noch ein Problem auf einer anderen Ebene.

Kann der CT allgemein als objektives Testverfahren gesehen werden, so wird durch die 'acceptable word'- Methode ein subjektiver Faktor eingeführt: ob eine Antwort akzeptabel ist oder nicht hängt von der subjektiven Beurteilung durch den Bewerter ab.

Alderson führte eine Sonderuntersuchung durch (1978:161-172), um die Zuverlässigkeit der Beurteiler zu überprüfen. Eine Retestmessung der Zuverlässigkeit der Native-beurteiler ergibt bei N=10 und n=30 Reliabilitätskoeffizienten (PEARSON-PRODUKT-MOMENT-KORRELATION) zwischen 0,93 und 0,98.

Eine Vergleichsgruppe von non-native speakers als Beurteiler ergibt eine durchschnittliche Retest-Zuverlässigkeit von 0,94. Ein Vergleich von Beurteilungen durch native speakers mit Beurteilungen durchgeführt durch non-native speakers ergibt Korrelationskoeffizienten zwischen 0,88 und 0,99.

Alderson folgert:

> "Non-native speakers agree in all essentials with native speakers as

to the acceptability of responses. Or rather if there is any disagreement it has little or no effect - the subjects' scores remain virtually the same.
The 'any acceptable' cloze scoring procedure is thus perfectly feasable in non-English speaking countries as well as in English-speaking countries."
Alderson (1978:170)

Anzahl der Aufgaben

Wie die Tabelle 16a zeigt, variiert die Menge der Testaufgaben in den Studien zur Fremdsprachenkompetenzmessung zwischen 25 (Pike, 1973) und 50 (z.B. Darnell, 1978; Oller, 1972; Hinofotis, 1976).
In den meisten Fällen scheint die Anzahl der Aufgaben mehr oder weniger intuitiv gewählt zu sein. Ein Blick in die Testtheorie zeigt, daß die Anzahl der Aufgaben nicht unwichtig ist, sondern einen direkten Einfluß auf die Genauigkeit hat, mit der ein Testinstrument messen kann. Im Kapitel über Reliabilität wurde bereits ausgesagt, daß die Reliabilität durch Verlängerung eines Tests, d.h. durch mehr Aufgaben, erhöht werden kann. Die Praktikabilität gebietet es jedoch, die Anzahl so weit wie möglich zu reduzieren. Wenige Untersuchungen machen Aussagen über die günstigste Anzahl von Aufgaben.
Haskell (1973) stellt fest, daß bei Testlängen zwischen 35 und 75 Aufgaben

"the number of deletion within this range did not effect the test results."
Haskell (1973:2-3)

Bormuth (1964) stellt fest - seine Untersuchung bezieht sich allerdings auf den muttersprachlichen Bereich -, daß CTs mit weniger als 50 Aufgaben dahin tendieren, unreliabel zu sein, da die Durchschnittswerte von Testformen der gleichen Textpassage aber mit unterschiedlicher Aufgabenzahl signifikant voneinander abweichen. Diese Aussage bezieht sich aber allerdings auf das Vertrauen, das man in die Gruppendurchschnittswerte setzen kann und nicht auf die von absoluten Werten unabhängige Rangordnung von Individuen.
Die einzige Studie, die das Problem der günstigsten Aufgabenzahl bei CTs ausführlich untersucht, stammt von Rand (1978). Gestützt auf Werte einer Untersuchung von James Brown an UCLA (1978) berechnet Rand den Zuverlässigkeitskoeffizienten für unterschiedliche Aufgabenmengen. Gleichzeitig

bezieht er in seine Berechnungen noch Unterschiede ein zwischen der 'exact scoring' und der 'acceptable word'- Methode. Das Ergebnis ist in einer Graphik der UCLA-Workpapers, June 1978, dargestellt.

ABB 5

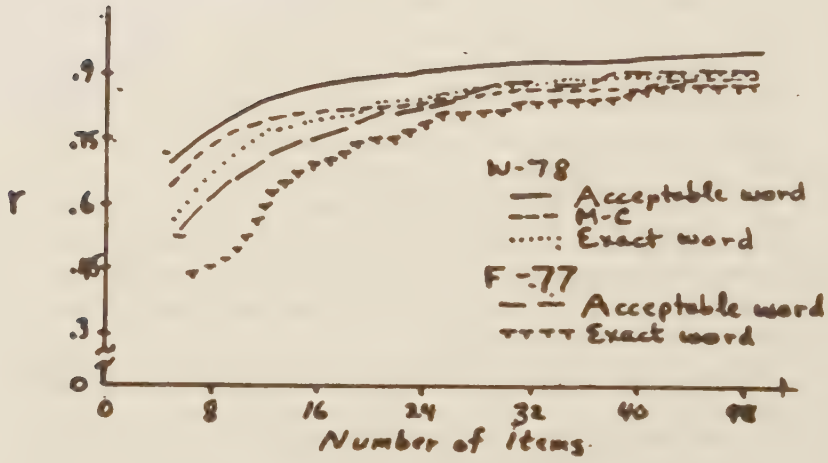


Figure 3

(entnommen aus: UCLA - Workpapers, June 1978:69)

Ab 25 Items flacht die Reliabilitätskurve für jede Auswertungsmethode so ab, daß ihr Anstieg bei mehr als 25 Aufgaben völlig unbeträchtlich ist. Rand faßt zusammen:

> "This study has shown that the examiner using cloze tests can most efficiently use his own resources and his examinees' time by giving a cloze test of only twenty-five deletions and employing the acceptable-word method of scoring."
> Rand (1978:o5,66)

Textschwierigkeit

Wie oben aufgezeigt (siehe z.B. Studie von Carroll, 1959; Darnell, 1968; Oller, 1972; Alderson, 1978) ist auch die Textschwierigkeit eine Variable des CT-Verfahrens.
Bevor der Einfluß dieser Variablen dargestellt wird, sollen kurz einige Erläuterungen zum Begriff der Textschwierigkeit folgen.
Die Erforschung und Messung der Schwierigkeit eines Textes ('Readability') hat im Gegensatz zur deutschen Forschung besonders in der amerikanischen Forschung eine umfangreiche Tradition. Während im Deutschen die Beurteilung der Textschwierigkeit und damit seine Angemessenheit für eine bestimmte Leser- bzw. Lerngruppe der Subjektivität eines Lehrers oder sonstigen Experten überlassen ist, wurden in der amerikanischen Forschung bereits 1939 von Lorge die ersten Formeln entwickelt, die eine objektive Beurteilung erlauben.[21]
Bekannte Formeln stammen von Flesch (1943), Dale & Chall (1948), Fog (1952), Spache (1953), Fry (1965), Bormuth (1969). Sie basieren meist auf einer Zählung bestimmter Textelemente, wie z.B. Silben, Wörter, Sätze, aber auch auf bestimmten Testformen. Als eine solche Testform ist ursprünglich auch der CT von Tylor entwickelt worden. Von den Ergebnissen, die eine Probandengruppe, z.B. Zehnjährige bei der Bearbeitung eines CT, erzielt, wird auf die Schwierigkeit des Textes in Relation zur Probandengruppe und gegebenenfalls zur Gesamtpopulation geschlossen. Beim CT, wie er hier untersucht wird, wird aus dem Ergebnis auf die sprachlichen Fähigkeiten der Probanden geschlossen. Es ergeben sich zwei Fragen:

1) Inwiefern variieren Individuen in ihren Reaktionen auf verschieden schwierige Texte,und inwiefern verändert sich der Zuverlässigkeitsgrad der Messung?
2) Ändern sich die Korrelationen mit den Validierungskriterien bei unterschiedlicher Textschwierigkeit?

Carroll et al. (1959) fanden, daß sich die Korrelation ihres CTs mit dem College Entrance Examination Board änderte, wenn sich die Schwierigkeit des CTs änderte

Darnell (1968) verglich vier CTs, zwei schwierige und zwei leichte, zwei über ein technisches Thema (engineering) und zwei über ein geisteswissenschaftliches Thema, mit dem TOEFL-Test, wobei er die Clozentropy-Auswer-

---

21 Eine umfassende Zusammenfassung ist zu finden bei Klare (1974-75), von wo auch diese Beispiele entnommen sind.

tungsmethode (siehe oben) verwendete. Die Schwierigkeit der Texte wurde nach der Formel von Flesch (siehe oben) bewertet. Es wurden zwar Unterschiede zwischen leichten und schwierigen Texten bei der Korrelation mit dem TOEFL festgestellt, aber sie waren minimal. Die Korrelationskoeffizienten betrugen beim leichten Text 0,69, beim schwierigen 0,63 im Bereich der technischen Thematik und beim leichten Text 0,73, beim schwierigen 0,77 im Bereich der geisteswissenschaftlichen Thematik.
Danach bestand der größere Unterschied in der Textthematik, denn die geisteswissenschaftlichen Texte korrelierten immer höher mit dem TOEFL als die technischen Texte (dabei waren die Probanden Studenten technischer Fächer). An sich überrascht das nicht unbedingt, wenn man einen motivationalen Faktor in Betracht zieht, der zwar nicht direkt greifbar, aber im sprachlichen Bereich von großer Bedeutung ist. Es ist leicht vorstellbar, daß der technische Text weniger motivierte, weshalb außenstehende Faktoren die Reliabilität negativ beeinflußten. Selbstverständlich kann eine solche Aussage, ohne den Text und die Probanden zu kennen, zu spekulativ sein.

Oller (1972) verwendete drei Texte, die alle leichter als Darnells Texte waren. Er bezeichnete sie "very easy", "fairly easy" und "standard" (mit Werten nach der Fleschskala von 100, 77 und 69). Dabei korrelierte der leichteste Text, Cloze I, konsistent beträchtlich niedriger mit dem Kriterium des UCLA-ESLPE-Tests (siehe auch S.77)als jeder der anderen zwei Tests. Die Korrelation mit dem Gesamttest (UCLA-ESLPE) betrug 0,73 für Cloze I, verglichen mit 0,87 für Cloze II und 0,85 für Cloze III. Zwischen Cloze II und III ist der Unterschied allerdings gering.

Alderson (1978:276) bezieht außer den Variablen der Textschwierigkeit (Easy / Medium / Difficult) auch die Lückenfrequenz (6, 8, 10, 12) und die Auswertungsmethode

a) exact scoring
b) any semantically acceptable word
c) any grammatically acceptable word
d) any word from the same form class as the deleted word
e) any word fulfilling the same grammatical function as the deleted word

in seine Untersuchung mit ein.
Bei den Korrelationskoeffizienten mit dem ELBA sind die besten Werte beim schwierigen CT nachzuweisen mit Ausnahme bei der Lückenfrequenz 6, wo der mittelschwere Text die besten Korrelationskoeffizienten ergab (siehe TAFEL 17).

TAFEL 17

(entnommen aus: Alderson, 1978:276)

Correlation of ELBA Total with various cloze tests

	TEXT		
	Easy	Medium	Difficult
letion Rate 6			
Exact	.59	.86	.81
SEMAC	.74	.88	.67
GRCO	.60	.81	NS
IDFC	.44	.67	.43
ACFC	.45	.68	.43
eletion Rate 8			
Exact	.70	.68	.82
SEMAC	.69	.77	.87
GRCO	.61	.74	.73
IDFC	.50	.51	.80
ACFC	.46	.50	.74
eletion Rate 10			
Exact	.65	.57	.79
SEMAC	.74	.74	.83
GRCO	.75	.75	.79
IDFC	.63	.70	.83
ACFC	.65	.65	.82
eletion Rate 12			
Exact	.67	.73	.77
SEMAC	.77	.78	.85
GRCO	.72	.75	.68
IDFC	.73	.70	.72
ACFC	.71	.69	.70

Dabei kann zusätzlich beobachtet werden, daß die 'any acceptable word'-Auswertungsmethode wiederum jeweils die höchsten Werte liefert (siehe T 16ab).
Alderson schließt, daß Unterschiede zwischen den bei CTs benutzten Texten in Bezug auf die Messung der Fremdsprachenkompetenz bestehen. Er vermag aber keine konsistenten Strukturen nachzuweisen.

Zusammenfassend kann diese Aussage im Hinblick auf die anderen dargestellten Forschungsergebnisse bestätigt werden.

L ü c k e n f r e q u e n z

Seit der "Erfindung" des CTs, der ja aus systematisch gebildeten Lücken innerhalb eines Textes besteht, wurden vor allem im muttersprachlichen Bereich aber auch im hier zu behandelnden fremdsprachlichen Bereich viele Möglichkeiten gefunden und erprobt, die Lücken zu setzen.

Alderson sieht verschiedene Interpretationsmöglichkeiten für die "systematische" Lückenbildung.
Eine Möglichkeit ist die Setzung der Lücken per Zufall ('random') oder Pseudozufall ('pseudo-random'), indem Wörter nach Zufallszahlen oder im Abstand bestimmter Frequenzen, z.B. jedes n-te Wort, ausgelassen werden.
Eine andere Interpretation von "systematisch" ist, daß die Wörter nach einem rationalen Prinzip ausgelassen werden, das in der Regel auf der Natur der Wörter selbst beruht. So können z.B. Wörter, die "leicht" oder "schwierig" zu finden sind, ausgelassen werden, oder solche, die als sehr redundant betrachtet werden.
Häufig ist die Methode, Wörter nach bestimmten Kategorien, wie z.B. Substantiva oder Verben, 'content' oder 'function words', auszulassen. Die Auslassung bestimmter grammatischer Formklassen scheint jedoch dem Test eine bestimmte Richtung zu geben. Dies wird z.B. in der Studie von Oller und Inal (1972) gezeigt, in deren CTs nur Präpositionen ausgelassen wurden. Die Korrelationen mit dem TOEFL als Validitätskriterium zeigten, daß die höchste Subtestkorrelation mit dem Grammatikteil bestand und alle anderen Subtestkorrelationen nicht signifikant waren. Die Autoren schlossen daraus, daß ihre Art des CTs ein gültiger Test grammatischer Kompetenz sei. Durch dieses Verfahren wird es offensichtlich möglich, die Clozetechnik auch für diagnostische Verfahren (siehe Kapitel I.1.1.) zu verwenden.

Beim Standard CT wird nur die Methode, jedes n-te Wort auszulassen, gebraucht. Es kann nicht übersehen werden, daß dieses Verfahren in gewissem Widerspruch zur traditionellen Testmethodik steht. Danach werden Aufgaben nur in einen Test aufgenommen, wenn sie bestimmte Anforderungen an die Trennschärfe und den Schwierigkeitsgrad erfüllen und voneinander unabhängig sind. Ein systematisches Testverfahren kann solchermaßen unakzeptable Aufgaben nicht vermeiden oder aussondern, sondern muß sie akzeptieren, da gerade diese Systematik ein wesentliches Charakteristikum des Tests ist.

Jonz (1976) bemerkt hierzu:

> "By using item analysis tc eliminate the less effective items in a cloze passage, one restores context. Thus, it can only be assumed by using item analysis, one might bolster one' s chances of reducing the number of deletions without irreparably undermining the test' s reliability."
> Jonz (1976:6-7)

Die systematische Methode der Itemauswahl verletzt möglicherweise auch noch die Forderung nach der Unabhängigkeit der einzelnen Aufgaben voneinander. Damit ist nicht ihre inhaltliche Homogenität oder Heterogenität gemeint, sondern inwieweit die Lösung eines Items von der Lösung des anderen abhängt. Sollte dies der Fall sein, ist die normale Anwendbarkeit der traditionellen Testtechnik anzuzweifeln.

Im Bereich des fremdsprachlichen CTs begann Carroll (1959) mit der Auslassung jedes 10ten Wortes, weil er jedes 5te für zu schwierig hielt. Auch Darnell verwendete die 10-Wort-Lückenfrequenz. Anderson (1970) ließ jedes 8te Wort aus und Stubbs & Tucker jedes 5te.
Die große Mehrzahl der Studien (Conrad, 1971; Oller, 1972; Irvine/Atai/Oller, 1974; Hinofotos, 1976; Krzyzanowski, 1976) hatten eine Lückenfrequenz von 7 Worten.
Haskell (1973) fand, daß Texte in der gleichen Reihenfolge eingestuft wurden ohne Unterschied, ob jedes 5te oder jedes 10te Wort ausgelassen wurde. Er fand auch keinen signifikanten Unterschied in den Durchschnittsergebnissen bei Lückenfrequenzen von 5, 7 oder 10 Wörtern.
Alderson präsentiert ebenfalls eine Untersuchung mit insgesamt 72 verschiedenen CTs. Im Vergleich der verschiedenen Variablen kommt er zur Bestätigung der oben gemachten Aussagen, daß unterschiedliche Lückenfrequenzen und somit unterschiedliche Kontextmengen keinen Einfluß haben auf die Vorhersagbarkeit der Aufgaben. Diese Aussage wird noch spezifiziert im

Zusammenhang mit dem Validitätskriterium:

> "The differing correlations of deletion rates with ELBA are not predictable. It is thus not possible to conclude that a more frequent deletion gives a d i f f e r e n t measure of English-as-a-foreign-language proficiency."
> Alderson (1978:290)

Nach diesen Aussagen bleibt also die Gültigkeit des Meßinstruments für die Messung der Fremdsprachenkompetenz unabhängig von der Lückenfrequenz. Alderson glaubt aber bei näherer Betrachtung der Ergebnisse einen anderen Schluß auf den Testinhalt machen zu können. Er stellt fest, daß verschiedene Lückenfrequenzen andere Testergebnisse ('rawscores') erbringen. Dies führt er nicht auf die unterschiedliche Kontextmenge, sondern auf unterschiedliche Lückenqualitäten zurück.

> "There is no evidence that the nature of the test changes, since validating correlations for different deletion rates do not differ, yet different means result for different cloze deletion rates. This can only mean that cloze scores are very much determined by w h i c h words have been deleted."
> Alderson (1978:370-371)

Daraus schließt er:

> "If cloze is so sensitive to the deletion of different words, then one wonders exactly what it is testing or indeed, whether it is capable of measuring higher order skills.
> A related point is that if cloze items, on the average, just as easy or as difficult with five words of context as with eleven words, then clozure must be based on a very small amount of context indeed. ... This provides evidence not only for the claim that cloze is sentence-bound, but goes even further, to suggest that cloze is clause or even phrase-bound. That being so, one would not expect it to be capable of measuring higher-order skills[21a], but rather to be a measure of much lower-order skills, and therefore be extremely sensitive to selection of different words, as indeed it has been seen to be."
> Alderson (1978:371) (Vergleiche auch Alderson, 1979:225 und 1980:74).

Bei diesem Schluß von statistischen Ergebnissen auf Operationsweisen des CTs liegt offensichtlich die Sichtweise zugrunde, daß der quantitative, durch den Lückenabstand erfaßte Kontext, und der zur Lösung des Items notwendige Kontext identisch sind. Diese Sichtweise ist aufgrund der Tatsache, daß es eine satzübergreifende Textkohäsion gibt, und daß selbst Wortkollokationen über Satzgrenzen hinaus bestehen können, anzuzweifeln.

Chihara, Oller, Weaver & Chavez-Oller (1977) testeten die Hypothese, ob cloze sentence-bound ist, an 201 japanischen Studenten der Universität

---

21a Alderson, 1978:372, bezeichnet als 'lower-order skills'
"the ability to handle syntax and lexis of the language (dubbed 'core proficiency') "
und er versteht unter 'high-order skills'
"the ability to make inferences and to handle the text".

von New Mexico mit verschiedenem Fremdsprachenkompetenzniveau Sie verwendeten zwei verschiedene Tests, von denen einer aus einer Textsequenz, der andere aus Einzelsätzen bestand. Sie wurden so an verschiedene Gruppen verteilt, daß vergleichbare Stichproben entstanden, mit den gleichen Texten einmal in Sequenz, das andere Mal in vertauschten Einzelsätzen. Die Ergebnisse zeigten, daß der Text in der natürlichen Sequenz bessere Ergebnisse lieferte als der Text aus vertauschten Einzelsätzen. Diese Unterschiede wurden auf höheren Kompetenzniveaus noch größer. Die Autoren schließen daraus, daß das Cloze-Verfahren nicht sentence-bound ist, sondern beeinflußt wird durch "discourse constraints ranging across sentences". Damit ist eine empirische Gegenaussage zu Aldersons Aussage gemacht. Sie beruht auf einer für die Überprüfung der Hypothese wesentlich sinnvolleren Methode, denn Chihara, Oller etc. konnten den Texteinfluß über den Lückenabstand hinaus nachweisen, was Alderson nicht gelang. Die Tatsache, daß die Ergebnisse von CTs bei verschiedenen Lückensetzungen differieren, wie es bei Alderson der Fall ist, dürfte eher im Repräsentationscharakter des 'samples' liegen, wie Alderson dies ja auch andeutet (1978:71; siehe Aldersons Schlußfolgerungen oben).

In der oben erwähnten methodisch-orientierten Studie von Porter (1978) wird die von Alderson (1978) gemachte Aussage unterstützt, daß verschiedene Auslassungen unterschiedliche Rohwertergebnisse bringen. Im Gegensatz zu Alderson schließt Porter nicht, daß CTs zwar das Konstrukt 'Language Proficiency' messen, aber auf andere Inhalte gerichtet sind, sondern, daß CTs nicht als zuverlässige Indikatoren einer allgemeinen Fremdsprachenkompetenz dienen können. Ein genauer Blick auf die Durchführungsmethode läßt jedoch an der Gültigkeit dieser Aussage zweifeln: CTs wurden durch Auslassung von insgesamt 50 Lücken im Abstand von acht Wörtern an einem literarischen (Test A) und einem nichtliterarischen Text (Test B) gebildet. Zum Vergleich wurden bei denselben Texten wiederum 50 Lücken gebildet, indem jeweils das folgende Wort ausgelassen wurde (Subtests A und B). Die 39 Probanden bearbeiteten die Tests im Abstand von drei Wochen. Porter bemerkt, daß der Übungseffekt, der sich für den zweiten Test ergibt, als irrelevant betrachtet wurde, da die gleiche Rangordnung der Studenten erwartet wurde. Hier scheint ein methodischer Fehler zu liegen, der an der Gültigkeit der Gesamtaussage zweifeln läßt:

a) Die Erwartung der gleichen Probandenrangordnung ist nicht nachgewiesen.

b) Da Test 1 die Antworten für Test 2 enthält, spielen bei der Durch-

führung von Test 2 Gedächtnisfaktoren eine überragende Rolle.

c) Die Vertrautheit mit dem Text, selbst wenn sie nur oberflächlich ist, ändert den Charakter des Testes wesentlich. (Siehe Beispiel S. 139).[21b] Insgesamt kann gesagt werden, daß die Begründung für Porters Zweifel an der Zuverlässigkeit (eigentlich müßte er an der Gültigkeit zweifeln) des CTs aus methodischen Gründen zurückgewiesen werden muß.

2.4.4. Theoretische Grundlagen des Cloze Tests

Aus dem bis jetzt Dargestellten wird ersichtlich, daß methodische Grundprobleme nicht losgelöst von theoretischen Grundfragen betrachtet werden können. Nimmt man z.B. die oben präsentierte These von Alderson, die er aus statistischen Ergebnissen abgeleitet hat, daß der CT im wesentlichen 'sentence-bound' ist, könnten aus einem CT ohne weiteres Sätze mit ineffektiven Aufgaben ausgelassen werden. Dadurch würde sich aber das gesamte Format des Tests radikal ändern. Im Endeffekt erhielte man unter Umständen einen Test aus losgelösten Einzelsätzen mit Lücken. Genau das ist jedoch der CT von seinem integrativ-pragmatischen Ansatz her nicht. Zur Rechtfertigung dieses Ansatzes ist es deshalb nötig, die Frage, was der CT mißt, zu untersuchen.

2.4.4.1. Empirische Antworten auf theoretische Grundfragen

Es gibt wenige empirische Antworten auf obige Frage. Der Nachweis von Korrelationen genügt hier nicht. Es kann wegen nachgewiesener Korrelationen keine qualitative Identität aus quantitativen Aussagen deduziert werden; in anderen Worten: wenn CTs sehr hoch mit anderen Tests korrelieren, heißt das nur, sie erbringen quantitativ das gleiche Ergebnis, müssen aber nicht notwendigerweise das gleiche messen (siehe auch das Kapitel 'Statistical Traps' in Oller, 1979:187 ff).
Ein Weg, empirisch zu erfassen, was der CT mißt, wurde von Weaver & Kingston (1963) mit einer faktorenanalytischen Studie begangen. Sie umfaßte insgesamt 18 Tests. 160 native speakers bearbeiteten den Davis Reading Test, verschiedene Subtests des Modern Language Aptitude Tests, Subtests der Ohio State Psychological Examination und einen Listening Subtest aus der 'Sequential Tests of Educational Progress' - Batterie. Dazu kamen acht CTs mit je 40 Aufgaben, von denen vier geschrieben, die vier verbleibenden auf Band aufgenommen wurden. Weaver & Kingston identifizierten mit

---

21b Auch hieraus ergeben sich Argumente gegen die Behauptung Aldersons (1978/1979/1980) CTs testen nur 'lower-order skills'.

Hilfe einer Varimax rotation drei Faktoren.

Faktor 1 hatte die größte Ladung bei den Reading Comprehension, Listening und Vocabulary Tests. Er wurde als 'verbal comprehension' bezeichnet.

Faktor 2 wurde als 'cloze factor' identifiziert, denn alle CTs hatten "Ladungen" von 0,49 bis 0,76. Dieser Faktor wurde als 'redundancy utilization' bezeichnet, denn die Lösungsmöglichkeiten der Aufgaben wurden zu einem großen Teil durch ihre kontextabhängige Voraussagbarkeit gegeben.

Faktor 3 wurde als 'rote memory' bezeichnet. Als Ergebnis stellen Weaver & Kingston u.a. fest,

> "there is much specific variance connected with the cloze procedure which is unexplained by the analysis."
> Weaver & Kingston (1963:261)

Auf der Basis der Studie von Weaver & Kingston argumentiert Carroll (1972), daß CTs eine Fähigkeit messen, die mit der Sprachkompetenz nicht verwandt ist. Er sagt vielmehr:

> "There is no clear evidence that cloze scores can measure the ability to comprehend or learn the major ideas or concepts that run through a discourse. ... cloze scores are probably more dependent on deletion of grammatical than of semantic cues. On the whole, the cloze technique in its usual form is too crude to permit measuring the degree to which the individual comprehends particular lexical or grammatical cues, or possesses a knowledge of specifical linguistic rules. It probably depends to a considerable extent on inferential processes."
> Carroll (1972:19)

Carroll drückt damit aus, daß die Schwäche des CTs darin besteht, daß er keine Aussage über die Sprachkompetenz machen kann, da er keine Einzelinformationen von Teilkompetenzen erfassen kann. Wenn man aber Sprachkompetenz so betrachtet, daß in ihr eine Anzahl von Teilkompetenzen integriert sind, dann ist das genau das, was Carroll als Schwäche des CTs sieht, seine Stärke. Sein Argument, daß CT-Ergebnisse nicht die Fähigkeit eines Probanden widerspiegeln, die Hauptgedanken eines Textes ("the major ideas or concepts that run through a discourse") zu erfassen,scheinen nicht völlig gerechtfertigt, wenngleich sie die Ansicht Aldersons (siehe oben), daß Cloze 'sentence-bound' wäre, bestätigen.

2.4.4.2. Betrachtungen der Funktionsweisen des Cloze Tests

Als Taylor den CT 1953 einführte, erklärte er dessen Funktionsweise durch die

> "human tendency to complete a familiar but not-quite-finished pattern to "see" a broken circle as a whole one, for example, by mentally closing the gap." Taylor (1953:415)

Sie wird auch als 'closure' bezeichnet. Ermöglicht wird diese Ansicht durch die Annahme einer holistischen Natur der Sprache. Dies bedeutet, daß Äußerungen als vollständige Einheiten produziert und wahrgenommen werden.

Um eine Lücke schließen zu können, sagt Taylor, bedarf der Proband des 'total language context'. Er definiert:

> "The total context of any language behavior includes everything that tends to motivate, guide, assist or hinder that behavior. It includes verbal factors - grammatical skills and multitudes of symbols - and non verbal ones such as fears, desires, past experience and intelligence."
> Taylor (1953:418)

Dies widerspricht der oben erwähnten Auffassung einer Sprachkompetenz von Lado, entspricht aber dem aufgestellten mehrschichtigen Kompetenzmodell.

Taylor erklärt die Wirkungsweise des 'total context of language':

> "I heard a _____ bark" is likely to elicit "dog", because that word is habitually associated with "bark" a n d because it fits in with past experience with noisy dogs. If the verbal context is enlarged to "For the first time, I heard a _____ bark" the impulse to supply "dog" may be reduced by common sense; the subject may ask himself "Who is this guy that has never heard a dog? Could he be referring to some other animal?" And if the preceding sentence has mentioned a voyage to the Pribilof Islands, the reader may draw on past knowledge to produce "seal"."
> Taylor (1953:418/419)

Solchermaßen werden auf "pragmatische" Weise Erklärungsmodelle wie Coserius 'lexikalische Solidaritäten' oder Leisis 'semantische Kongruenzen' für syntagmatische Beziehungen in einem Text erweitert. Neben der Fähigkeit zur Closure dient die Fähigkeit eines Sprechers, die R e d u n d a n z eines Textes zu nützen, zur Lösung von CTs.

Taylor verwendet eine Erklärung von Osgood (1952 und 1953):

> " "Man coming" means the same as "A man is coming this way now". The latter which is more like ordinary English, is redundant; it indicates the singular number of the subject three times (by "a", "man" and "is"),the present tense twice ("is coming" and "now"), and the direction of action twice ("coming" and "this way"). Such

> repetitions of meaning, such internal ties between words, make it possible to replace "is", "this", or "now", should any of them be missed."
> Taylor (1953:418)

Taylor muß hier insofern noch ergänzt werden, als redundante Elemente vom Kontext beeinflußt und nur durch dessen Kenntnis richtig erkennbar sind. Ist im Kontext des oben genannten Beispiels z.B. schon die Rede gewesen von einem b e s t i m m t e n Mann, so ist nicht "a", sondern nur "the" als akzeptables fehlendes Element zu ersetzen; dies gilt natürlich in ähnlicher Weise für die anderen Elemente dieses Beispiels.
Sprachliche Äußerungen sind also zu einem gewissen Grad redundant. Das heißt, es gibt Elemente in einer Äußerung, die zum Dekodieren ihres Informationsinhaltes nicht notwendig sind, auch wenn sie noch andere Informationen über die Äußerung beinhalten. Redundante Elemente werden trotzdem im Kommunikationsprozeß beibehalten, denn sie vermindern die Irrtumswahrscheinlichkeit und verhindern Mißverständnisse. Sie werden auch beibehalten, um die Anforderungen eines bestimmten Stiles zu erfüllen, und ihre Anwesenheit oder ihr Fehlen kann dem kompetenten Sprecher Auskunft geben über das Kompetenzniveau des Senders. So ist z.B. Kindersprache oder Ausländersprache leicht an fehlenden redundanten Elementen zu erkennen.
Darüber hinaus ermöglichen redundante Elemente auch ein Gelingen der Kommunikation, selbst wenn der Übertragungsvorgang gestört wird. Die Redundanz kann aber nur bis zu dem Grad benützt werden, wie der Empfänger kompetent ist. Leicht beobachtbar ist das Phänomen, daß weniger kompetente Sprecher mehr redundante Elemente benötigen. Lernende verlangen z.B. häufig nach einem langsameren Sprechtempo oder nach größerer Lautstärke, um alle redundanten Elemente für ihren Dekodierungsprozeß verwenden zu können. Die Anzahl benötigter redundanter Elemente und die Sprachkompetenz verhalten sich also umgekehrt proportional. Diese Tatsache nützt der CT. Beim CT, so wird argumentiert, wird die Notwendigkeit redundanter Elemente, die durch Zufall bestimmt sind, und dadurch gewissermaßen repräsentativ sein sollen, quantitativ erfaßt und von der Anzahl erkannter Elemente auf die Sprachkompetenz geschlossen.
Es erhebt sich die Frage, ob die Redundanz auf allen Kompetenzstufen wirkt. Im Rahmen dieser Arbeit interessiert die Frage, ob sie auch auf Stufen niedriger Kompetenz wirkt, bzw. ob und mit welcher Qualität der CT auch in Anfangsklassen des Fremdsprachenunterrichts eingesetzt werden kann.

Eine weitere Möglichkeit, die Funktionsweise des CT zu erklären, bilden die sog. 'Transitional Probabilities'. Taylor erklärt:

> "Some words are more likely than others to appear in certain patterns or sequences. "Merry Christmas" is a more probable combination than "Merry birthday". "Please pass the _____" is more often completed by "salt" than by "sodiumchloride" or "blowtorch". Some transitions from one word to the next are, therefore, more probable than others." Taylor (1953:418)

Dies entspricht auch den Ergebnissen der sog. Kollokationsforschung, wo ebenfalls von einer Reihe verschieden starker syntagmatischer Bindungen, die sich auf einem Kontinuum zwischen zufälliger und fester Kollokation befinden, gesprochen wird.

Deutlich wird auch hier wieder die dominierende Kraft des Kontextes: selbst wenn normalerweise von einer größeren Wahrscheinlichkeit gesprochen werden kann, daß im Beispiel oben eher "salt" als "blowtorch" einzusetzen ist, ist dies hinfällig, wenn im Kontext von einigen Schweißern bei der Arbeit gesprochen wird.

Sichtbar wird in diesem Zusammenhang auch die Schwäche eines Erklärungsmodells, das im Sinne Chomskys mit 'Selektionsbeschränkungen' operiert und grundsätzlich nur satzorientiert ist.

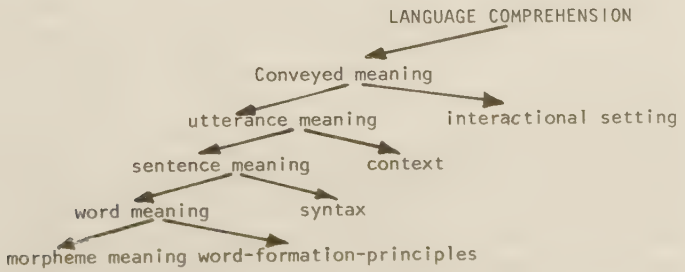
Lassen wir die bisher behandelten Aspekte der Vollständigkeit der Information, die Redundanz eines Textes, die 'Transitional Probabilities' außer acht, so ist die Grundlage für das Zustandekommen eines erfolgreichen Kommunikationsaktes die Kommunalität der Erfahrungshintergründe der Sprachpartner. Taylor hat diese Kommunalität nach Osgood "corresponding sets of habits", einem behavioristischen Sprachmodell entsprechend, bezeichnet.

In der Annahme, daß Kommunikationsprozesse auch kognitive Prozesse erfordren, soll eine behavioristische Einengung des Begriffs der Kommunalität vermieden und hier allgemein von einer Kommunalität des Erfahrungshintergrundes gesprochen werden. Dieser Erfahrungshintergrund ist nicht nur der sprachliche, sondern auch der gesamte außersprachliche Erfahrungsbereich. Besteht hier keine gegenseitige Vertrautheit der Partner bricht die Kommunikation zusammen. Bleiben wir beim obigen Beispiel "Please pass the _____": Nehmen wir an, das kontextuell akzeptable, wenn auch von der 'transitional probability' her seltenere Wort ist 'sodiumchloride'; die Kommunikation bricht zusammen, wenn der Empfänger mit dem Sender nicht das Wissen um den Begriff teilt, es sei denn, es gelingt dem Sender, den Erfahrungshorizont des Empfängers durch entsprechende Erklärungen zu erweitern.

Mit den dargestellten Begriffen 'Closure', 'Total Context of Language', 'Redundanz', 'Transitional Probabilities', 'Kommunalität von Sprachbenutzern' wird die Funktionsweise des CT aus der Sichtweise verschiedener Disziplinen erklärt: Gestaltpsychologie, Sozio- und Pragmalinguistik, Syntax und Semantik als Bereiche der allgemeinen Linguistik und Kommunikationswissenschaft.

Ein weiteres Erklärungsmodell für die Funktionsweise des CT liefert Oller (siehe auch Kapitel 2.2.) aus psycholinguistischer Sicht mit der Darstellung der sog. 'Expectancy Grammar'.[21c]
Auf der Basis des 'Analysis-by-synthesis'- Modells stellt der Sprachbenutzer sowohl bei rezeptiven als auch bei produktiven Sprachleistungen bezugnehmend auf den Kontext konstant antizipierende Hypothesen zum Textverlauf auf.

All diese Erklärungen zur Funtkionsweise sind theoretischer Art. Eine sichere Aussage ist nur die, daß der CT die "Clozefähigkeit" mißt, d.h. die Fähigkeit zum Schließen von Lücken in einem unvollständigen Diskurs / Text. Die Darstellungen haben aber gezeigt, daß diese "Clozefähigkeit" differenziert gesehen werden muß. Sie ist kontextbezogen und operiert auf allen sprachlichen Ebenen. Dabei muß der Proband sowohl rezeptive Leistungen erbringen, um den Kontext zu verstehen, als auch produktive Leistungen, um die Lücken zu füllen.
Er muß Morphembedeutungen, Wortbedeutungen, Satzbedeutungen, Utterancebedeutungen und die Gesamtbedeutung des präsentierten Textes erkennnen und innerhalb von Prinzipien der Wortbildung, der Syntax, des Kontextes und des "interactional setting' in akzeptabler Weise reagieren. Dies entspricht dem Modell der Language Comprehension von Fillmore[22] (1979):



21c Vgl. auch Kenneth Goodman (1967 und 1970) über das Phänomen der 'Expectancy Grammar' in "Some Sources of Reading Problems for Foreign-language Learners", C.A. Yorio (Univ. of Michigan), in: Language Learning, Vol. 21, No.1, June 1971, und H.L. Jorstad, "Testing Communications" in: Jarvis, 1974.

22 Präsentiert von Fillmore anläßlich eines Gastvortrages an der Universität München am 19. Juli 1979

Der CT wird in seinem integrativen Charakter bestätigt: er aktiviert die linguistischen Teilkompetenzen der Code Competence und die Communicative Competence gleichzeitig. Im Bereich der Communicative Competence wird der rezeptive Teilbereich durch den Vorgang des Kontextverstehens direkt aktiviert. Es ist fraglich, inwieweit der produktive Vorgang der Lückenschließung dem produktiven Bereich der Code Competence mit der dialogischen Teilkompetenz, der mündlichen und der schriftlichen Teilkompetenz entspricht.
Akzeptiert man Ollers Modell der 'Expectancy Grammar', besteht eine Entsprechung auf einer tieferen Ebene, durch die allen sprachlichen Vorgängen zugrundeliegende 'Expectancy Grammar'. Man kann auch fragen, ob dieser Teilbereich gegebenenfalls nicht indirekt erfaßt werden kann durch entsprechende Textinhalte. So kann ein Dialog, ein mündlicher oder schriftlicher Bericht zum Inhalt eines CTs gemacht werden, und die Reaktion des Probanden darauf in indirekter Weise seine dialogische oder schriftliche Teilkompetenz wiederspiegeln. Andeutungsweise enthielten die untersuchten Tests zwar verschiedene Textsorten, spezifische Untersuchungen stehen aber aus.
Insgesamt kann aus den theoretischen Überlegungen geschlossen werden, daß der CT dem aufgestellten Modell der Fremdsprachenkompetenz entspricht. Damit befindet er sich auch im Einklang mit der Zielsetzung des schulischen Unterrichts, dem dieses Kompetenzmodell zugrunde liegt.

Z u s a m m e n f a s s u n g :

Nach einer Skizze der Entwicklung vom Readability-Test für native speaker (Taylor, 1953) zum weltweit verwendeten Standardmodell, das nach Oller, 1979, definiert wird, wurden Gründe für die Attraktivität des CT präsentiert. Dies sind besonders eine P r a k t i k a b i l i t ä t - insbesondere gegenüber traditionellen Test - und eine psychologische V a l i d i t ä t , die jedoch nicht unbestritten ist (z.B. Clark, 1978). Ein Überblick über die bisher geleistete empirische Forschung zum Standard CT als Meßinstrument der Fremdsprachenkompetenz zeigt, daß der CT im allgemeinen ein r e l i a b l e s T e s t v e r f a h r e n ist. Die Validitätskoeffizienten bewegen sich zwischen 0,41 und 0,89, die Reliabilitätskoeffizienten zwischen 0,61 und 0,95. Die Validität wird in der Regel durch Korrelierung mit als g ü l t ig definierten Meßinstrumenten der Fremdsprachenkompetenz nachgewiesen. Als solche Validierungskriterien dienen etablierte Testverfahren und Expertenurteile. Erhebliche Unterschiede zeigen die Untersuchungen bei den Korrelationsergebnissen mit den Subtests der Validierungskriterien. Weder die Behauptung, der CT kor-

reliere eher mit integrativen Tests noch mit 'discrete-point'- Tests findet allgemeine Bestätigung.

Als kontrollierbare Variable des CT werden Auswertungsverfahren, Lückenfrequenz, Aufgabenzahl, Testlänge und Testschwierigkeit erkannt. Im allgemeinen werden die besten Ergebnisse für das 'acceptable word'-Auswertungsverfahren, eine Lückenfrequenz zwischen 5 und 10 Wörtern und eine Aufgabenzahl von 50 Items erreicht. Rand weist jedoch eine größere Praktikabilität und Effizienz bei 25 Testitems nach. Schwierige Texte scheinen dahin zu tendieren, bessere Ergebnisse zu liefern.
Die meisten Autoren schließen aus den erhaltenen Ergebnissen, daß der CT ein gültiges Meßinstrument einer allgemeinen Fremdsprachenkompetenz sei, was aber nicht gänzlich unwidersprochen bleibt. Alderson (1978) behauptet, daß der CT keine Kontextzusammenhänge erfassen könne. Dieser These stehen eine Reihe semitheoretischer Überlegungen gegenüber.
Porter zweifelt, daß der CT ein gutes Meßinstrument einer allgemeinen Fremdsprachenkompetenz sei (1978). Aufgrund methodischer Fehler bei seinen Untersuchungen müssen diese Zweifel jedoch zurückgewiesen werden.

Es ist ein allgemeines Manko der vorliegenden Abhandlungen und Untersuchungen, daß die Reliabilitäts- und Validitätskriteriennachweise wegen der Verschiedenartigkeit der Variablen und der Probandengruppen kaum verglichen werden können.

Unklare empirische Aussagen bestehen auch darüber, was der CT messe. Eine Übersicht theoretischer Betrachtungen ergibt, daß der CT keine einheitliche Skill mißt, sondern auf mehreren Grundlagen beruht: auf der auf Gestaltprinzipien beruhenden Fähigkeit zur 'Closure', der Fähigkeit des kompetenten Sprechers, die Textredundanz zu erkennen und zu nutzen, der sog. Kommunalität der Kommunikationspartner und der sog. 'Expectancy Grammar' von Oller.

Der CT wird in seinem integrativen Charakter und in seiner Kongruenz mit dem aufgestellten Modell der Fremdsprachenkompetenz, das dem Englischunterricht zugrunde liegt, bestätigt.

3. DISKUSSION UND PROBLEMATIK

Im Anschluß an diesen Überblick über die vorhandene Literatur und die Grundlagenforschung sollen die wesentlichen Punkte noch einmal zusammengestellt werden, und die daraus erwachsende Problematik für die vorliegende Untersuchung aufgezeigt werden.

Eine Reihe von Forschungsfragen bestimmen Ziel und Richtung der Studie. Die erste Forschungsfrage nach der Brauchbarkeit des CT erfordert eine Übersicht über die Intentionalität von Fremdsprachentests: Zur Beantwortung der Frage, wozu Fremdsprachentests gebraucht werden. Aus der in der Literatur vorgefundener terminologischen Vielfalt ist erkennbar, daß Fremdsprachentests in unterschiedlichen Funktionen und Formen für verschiedene Zwecke und Zielgruppen verwendet werden.

In Abgrenzung zu 'discrete-point' Tests, die einzelne Teile der Fremdsprachenkompetenz messen, wurde gezeigt, daß der CT ein integrativer Test ist, der als Meßinstrument einer allgemeinen Fremdsprachenkompetenz dienen kann. Somit mißt er nicht Stärken und Schwächen einzelner Teilkompetenzen der Lernenden, sondern er soll eine Beurteilung der allgemeinen Fremdsprachenkompetenz erbringen. Die Entscheidung für oder gegen die eine oder andere Testart ist keine Entweder/Oder-Entscheidung (disjunctive fallacy), sondern abhängig von der beabsichtigten Meßfunktion.

Die Fremdsprachenkompetenz wurde als integratives Konstrukt der sog. Code Competence und der Communicative Competence dargestellt. Die Code Competence besteht aus der phonologischen, lexikalischen, grammatikalischen und semantischen Teilkompetenzenz, mit den rezeptiven und produktiven Teilbereichen bestehend aus den dialogischen, mündlichen und schriftlichen Teilkompetenzen.

Überlegungen zu den Operationsweisen des CT haben gezeigt, daß der CT als integrativer Test dem mehrschichtigen Kompetenzmodell, das dem fremdsprachlichen Unterricht zugrunde liegt, entspricht. Im einzelnen wurden die Fähigkeit zur 'Closure', die Einbeziehung des 'Total Context of Language', die Redundanz, die 'Transitional Probabilities', die Kommunalität der Sprachbenutzer und die 'Expectancy Grammar' als Grundlagen des Clozeverfahrens dargestellt.

Seine Verwendbarkeit für bestimmte administrative Zwecke, z.B. Kurszuteilung oder Einstufung der Probanden innerhalb einer Lerngruppe hängt von seiner Qualität ab. Als Qualitätskriterien von Tests wurden

Objektivität
Reliabilität
Validität
Praktikabilität und
Normierung

dargestellt, wobei die Objektivität gesehen werden kann als

Durchführungsobjektivität
Auswertungsobjektivität und
Interpretationsobjektivität.

Es kann wohl mit Sicherheit gesagt werden, daß der CT von seiner Anlage her einen hohen Grad an Durchführungsobjektivität, d.h. Unabhängigkeit von Testadministratoren erwarten läßt. Für die Durchführung des Experiments ist für objektive Bedingungen zu sorgen. Entsprechende Kriterien wurden aufgestellt.

Die Auswertungsobjektivität hängt von der verwendeten Methode ab. Den höchsten Objektivitätsgrad bringt die sog. 'exact scoring'- Methode. Probleme bestehen bei der 'acceptable word'- Methode, da die Akzeptabilität von der Subjektivität des Beurteilers abhängt. Eine Studie von Alderson (1978) hat jedoch gezeigt, daß sowohl native als auch non native Beurteiler durchaus zuverlässige Ergebnisse liefern können.

Die Grundlage für die Interpretationsobjektivität ist durch die numerisch eindeutigen Testergebnisse ('raw scores') klar gegeben.

Zur Bestimmung der Reliabilität wurden vier Methoden präsentiert:

Retestreliabilität
Paralleltestreliabilität
Split-half Reliabilität und
Konsistenzanalyse.

Die Retestreliabilität erfordert eine zweimalige Durchführung des gleichen Tests. Diese verbietet sich beim CT, da seine erstmalige Lösung die Aufgabenstellung für den zweiten Bearbeitungsgang grundsätzlich ändern würde. Gedächtnisfaktoren würden eine systemändernde Rolle spielen. Solange die Abhängigkeit des CT vom Textinhalt noch nicht völlig erforscht und anlysiert ist, ist es kaum möglich, systematisch parallele Testformen zu entwickeln, die für eine Feststellung der Paralleltestreliabilität nötig sind.

Die meisten der vorliegenden Studien berechnen deshalb die als 'Innere Konsistenz' bezeichnete Art von Reliabilität. Als Verfahren werden die

sog. Split-half Methode oder die Konsistenzanalyse verwendet. Die vorliegenden Studien zeigen, daß recht brauchbare Werte erzielt wurden, sodaß der CT allgemein als reliabler Test gesehen wird. Bei der Interpretation der Reliabilitätskoeffizienten sind eine Reihe möglicher Störfaktoren zu beachten:

Testhomogenität
irrepräsentative Variabilität
Testlänge
Streuungsbereich der Aufgabenschwierigkeit und
Aufgabeninterkorrelationen.

Eine A u f g a b e n a n a l y s e und die k o n s e q u e n t e A u s s o n d e r u n g ungeeigneter Aufgaben, die die Anforderungen der Trennschärfe nicht erfüllen, ist beim Standardmodell des CT mit einer systematischen Aufgabenauswahl nicht möglich.

Eine weitere Problematik besteht in der Frage der A u f g a b e n i n t e r k o r r e l a t i o n e n . Aufgrund der Textkohärenz und der systematischen Aufgabenauswahl innerhalb des fortlaufenden Textes ist es fraglich, ob die Aufgaben im Sinne der traditionellen Testtheorie als unabhängig voneinander betrachtet werden können.

Es wurden vier Arten der V a l i d i t ä t vorgestellt:

Inhaltliche Validität	(content validity)
psychologische Validität	(face validity)
Konstruktvalidität	(construct validity)
kriteriumsbezogene Validität.	(criterian referenced validity).

Als wichtigste wird von Lienert die kriteriumsbezogene Validität gesehen, weil sie empirisch erfaßbar und quantifizierbar ist. Dabei wird ein Kriterium, von dem man annimmt, daß es in Bezug auf das zu messende Persönlichkeitsmerkmal Gültigkeit besitzt, mit dem zu untersuchenden Test korreliert. Wenn ein an sich valides Kriterium, z.B. ein Expertenurteil verwendet wird, spricht man von ä u ß e r e r V a l i d i t ä t , wird ein Kriterium verwendet, das selbst bereits validiert ist, z.B. ein standardisierter Test, spricht man von i n n e r e r V a l i d i t ä t, wobei die erstere wegen ihrer größeren Direktheit vorzuziehen ist.
In der dargestellten Literatur wurden CTs sowohl wegen ihrer äußeren als auch inneren Validität als valide Tests der Fremdsprachenkompetenz dargestellt. Die Nachweise der inneren Validität erscheinen problematisch. Als Validierungskriterien dienten in der Regel standardisierte Tests (TOEFL etc), die auf der Basis des 'discrete-point approach' aufgebaut sind. Wenngleich im allgemeinen hinreichende

Kriterien für die Gesamtvalidität festgestellt wurden, ergaben sich unterschiedliche und zum Teil widersprüchliche Ergebnisse in Bezug auf die Subtestkorrelationen. Es erscheint deshalb fraglich, ob ein mit einer ganz anderen Funktion entworfener Test ein zweckmäßiges Validierungskriterium sein kann. Hinzu kommt die Tatsache, daß die Reliabilitätskoeffizienten des Validitätskriteriums möglicherweise einen Einfluß auf die Validitätskoeffizienten haben, also höchste Ansprüche erfüllen müssen, was nicht immer - vor allem bei Subtests - gewährleistet ist.
Für den deutschen Bereich liegen keine standardisierten Tests vor, die diese Reliabilitäts- und Validitätsansprüche erfüllen können.
Die günstigste Lösung liegt daher in einer Verwendung des Lehrerurteils als Validierungskriterium, selbst auf die Gefahr dessen möglichen Subjektivität hin. Ingram bestätigte jedoch die Zuverlässigkeit des Lehrerurteils. (Ingram, 1977:21)
Die übrigen Arten der Validität können nicht empirisch untersucht werden und quantitativ festgelegt werden, sondern sie müssen auf theoretischer Basis eruiert werden.

Die dargestellten theoretischen Betrachtungen zeigten, daß plausible postitive Aussagen zur Content-, Konstrukt- und psychologischen Validierung des CT gemacht werden können.
Der CT entspricht in seiner Funktionsweise normalen kommunikativen Prozessen. Der CT-Bearbeiter muß zunächst den empfangenen Text dekodieren,[22a] die Information verarbeiten und produktiv darauf reagieren. Sollen diese Fähigkeiten erfaßt werden, besitzt der CT Content Validity.
Sieht man den Gebrauch der sog. 'Expectancy Grammar', der Redundanz, der 'Transitional Probabilities', der Kommunalität der Sprachbenutzer als Teile des Konstrukts der Fremdsprachenkompetenz, besitzt der CT Konstruktvalidität.

In ihrer Darstellung wurden diese Anrahmen zu den verschiedenen Validitätsarten als plausibel erkannt. Diese Plausibilität kann zwar nicht direkt quantitativ erfaßt werden, aber sie kann durch entsprechende empirische Nachweise an Gewicht gewinnen.

Das Gütekriterium der P r a k t i k a b i l i t ä t des Tests ist im allgemeinen als die offensichtlichste und überzeugendste Qualität des CT erkannt worden. Für die Durchführung der Untersuchung wurde ein detail-

---

22a Dabei setzen hier bereits produktive 'anticipatory strategies' ein (vgl. 2.4.4.2.).

lierter Katalog von Merkmalen seiner Bestimmung dargelegt.

Auch für die Normierung der Tests soll die unternommene Studie Grundlagen schaffen und die Fragen klären, ob der CT die Ansprüche, die die Normierung stellt, erfüllen kann und .. ob er als informeller Klassentest oder als einfache, vom Lehrer selbst zu erstellende Aufgabe im Unterricht verwendbar ist.

Die vorliegenden Untersuchungen ließen eine Reihe von Variablen des Cloze Tests erkennen, die seine Qualität beeinflussen:

Auswertungsmethode
Lückenfrequenz
Aufgabenanzahl
Textschwierigkeit.

Unter den verschiedenen Auswertungsmethoden wurde übereinstimmend in allen Untersuchungen festgestellt, daß die 'acceptable word'- Methode bessere Reliabilitätswerte und Validitätskoeffizienten erbringt. Der Nachteil der größeren Subjektivität dieser Methode gegenüber der 'exact scoring'- Methode erscheint in Anbetracht der nachgewiesenen Zuverlässigkeit des Lehrerurteils und großer pädagogischer Vorteile vernachlässigenswert. Für die vorliegende Untersuchung wurde deshalb die Entscheidung zugunsten der 'acceptable word'- Methode getroffen.

Als Lückenfrequenz wurde für den Standard CT allgemein ein Wortabstand zwischen 5 und 10 Worten festgelegt. Da die vorliegenden empirischen Untersuchungen zeigten, daß innerhalb dieses Spielraums kaum wesentliche Unterschiede der Ergebnisse auftraten, wird hier die im non-native Bereich häufigste Frequenz von 7 Wörtern angewandt.

Obwohl die Aufgabenzahl von 50 Items als in der Mehrzahl der vorliegenden Literatur und Untersuchungen üblich bezeichnet wird, geben zwei Gründe den Ausschlag dafür, nur 25 Aufgaben pro Test zu verwenden:

1) Die Untersuchung von Rand (1979) besagt, daß 25 Aufgaben die effektivsten Ergebnisse erbringen, und daß darüber hinaus keine wesentliche Steigerung der Testqualität zu erwarten ist. (1979 : 366)
2) Außerdem haben Versuche die Feststellung Ollers verifiziert, daß pro Item mit etwa einer Minute Arbeitszeit gerechnet werden muß. Zählt man Testanweisung, Einsammeln und Austeilen sowie genügend Zeit auch für langsame Schüler, mit denen vor allem auf unteren Niveaustufen

gerechnet werden muß, so füllt ein aus 25 Items bestehender Test genau eine 45-minütige Unterrichtsstunde.

Die Variable der T e x t s c h w i e r i g k e i t ist von besonderer Bedeutung, da der Text sozusagen das "Grundmaterial" des Tests liefert. Zur Bestimmung der Schwierigkeit englischsprachiger Texte liegen Formeln vor. Die Hauptproblematik besteht in der Zuordnung des richtigen und angemessenen Schwierigkeitsgrades für die Kompetenzstufe einer Probandengruppe. Forschungsergebnisse zeigen eine Tendenz dahingehend, daß schwierige Texte bessere Reliabilitäts- und Validitätsergebnisse liefern. Hier müsser noch Orientierungshilfen zur Bestimmung zu leichter und allzu schwieriger Texte gefunden werden.

Neben diesen methodischen Variablen ist die wichtigste Variable die der Probanden. Die Population, für die der Test untersucht werden soll, sind Schüler der Klassen 5 - 9 unseres dreigliedrigen Schulsystems, das charakterisiert wurde durch eine typische Leistungsstruktur der Fremdsprachenkompetenz (vgl. 2.3.).

a) Es besteht ein Leistungsgefälle von den höheren zu den niederer Klassen der jeweiligen Schulart.

b) Die Klassen selbst sind relativ homogene Lerngruppen.

c) Es besteht ein Leistungsgefälle vom Gymnasium über die Real- zur Hauptschule.

d) Die ausgewählten Klassen sind Lerngruppen des elementaren Fremdsprachenunterrichts, weshalb die Kompetenzstufen relativ niedrig sind.

Für die Brauchbarkeit des CTs bei Probandengruppen dieser Art liegen bislang keine Untersuchungen vor. Es sind auch noch keine Untersuchungen verfügbar für Probanden mit Deutsch als Muttersprache und für die Anwendbarkeit auf niedrigen Kompetenzstufen der verschiedenen Schularten. Ergebnisse von anderen Probandengruppen sind nicht einfach übertragbar, da nachgewiesenermaßen sowohl der muttersprachliche Hintergrund der Probanden als auch ihre Kompetenzstufen die Validität und Reliabilität beeinflussen.

4. METHODE

Es ist das Ziel der vorliegenden Arbeit, den CT auf seine Brauchbarkeit in unserem Schulsystem zu prüfen.

In Kapitel I wurden die Notwendigkeit und die Zielrichtung der Studie dargelegt und vorgestellt.

Kapitel II präsentierte eine Übersicht über theoretische Grundlagen und die bereits durchgeführte Forschungsarbeit.

In Kapitel III wurden die wesentlichen Ergebnisse zusammengefaßt und die daraus erwachsende Problematik dargestellt.

Kapitel IV beschreibt die praktische Durchführung des Versuchs, der zur empirischen Beantwortung der Forschungsfragen dienen soll. Das Kapitel beginnt mit der Darstellung der Vorversuche und ihrer Ergebnisse, beschreibt die Hypothesenbildung, die Planung für die Hauptuntersuchung und die Analyse der Probandengruppe, die Konstruktion des Tests sowie die praktische Durchführung des Tests.

4.1. Vorversuche

Veranlaßt durch Berichte in der Fachliteratur, jedoch ohne Vorbilder aus dem deutschsprachigen Raum, ganz zu schweigen von Vorbildern aus unserem Schulwesen, wurden in der schulpraktischen Arbeit des Lehrstuhles für die Didaktik der englischen Sprache und Literatur an der Universität München die ersten exploratorischen Versuche mit dem CT durchgeführt.

Erste informelle Fragestellungen waren dabei, ob der Test überhaupt bei uns durchführbar ist, in welcher Form dies geschehen soll, und welche Ergebnisse er bringt.
Als Zielgruppe dienten zunächst Klassen aller Niveaustufen und Schularten, die zur Verfügung standen.[23]
Die Ergebnisse waren recht ermutigend. Die Tests bereiteten kaum Schwierigkeiten bei der Durchführung, sie wurden von den Schülern im allgemeinen positiv bewertet und aufgenommen, und rein intuitiv schienen die Ergebnisse mit den bekannten Fähigkeiten der Schüler übereinzustimmen.
Als numerisch faßbarer Ausdruck für die Fremdsprachenkompetenz wurde die jeweils letzte Zeugnisnote verwendet. Es wurde angenommen, daß sie die

23 Die Genehmigung des Bayerischen Staatsministeriums für Unterricht und Kultus liegt vor.

Fremdsprachenkompetenz am besten repräsentiert, da sie das Gesamtergebnis einer Batterie von Prüfungsaufgaben darstellt, die sich über das ganze Schuljahr verteilen. Von den Prüfungsaufgaben wird angenommen, daß sie alle fremdsprachlichen Teilbereiche repräsentieren, indem sie mündliche und schriftliche Leistungen aller möglichen Gebiete erfassen. Darüber hinaus werden durch die Vielzahl der einzelnen Leistungen auch Leistungsschwankungen ausgeglichen und ein "Normalpegel" festgesetzt. Gelegentlich wichen die CT-Ergebnisse von den Schulnoten ab. Hier kam die erste Überraschung: als die Lehrer dazu befragt wurden, korrigierten sie in den meisten Fällen den Aussagewert der Schulnote. Sie stellten fest, daß in die Zeugnisnoten Faktoren eingeflossen waren, die für die tatsächliche Fremdsprachenkompetenz irrelevant sind; z.B.:

Schüler erhielten wegen Unterschleifs die Note ungenügend. Ihre Zeugniszensur war deshalb schlechter als sie tatsächliche Leistung.

Schüler erbrachten viele reine Lernleistungen, was sich insbesondere bei sog. "Kurzarbeiten" oder ähnlicher Arbeiten ausdrückte, bei denen nur der Lernstoff - etwa die Vokabelkenntnis - abgefragt wurde. Aufgrund konstanter Lernleistungen waren deshalb die Zensuren besser als die "tatsächliche " Fremdsprachenkompetenz der Schüler.

Schüler konnten wegen Krankheit bestimmte Lernstoffe nicht bewältigen und deshalb einen bestimmten Leistungsstand nicht erreichen.

Schüler waren emotional gestört und konnten deshalb ihre tatsächlichen Leistungen nicht erbringen.

Gelegentlich erkannten Lehrer, daß auch die CT-Ergebnisse von irrelevanten Faktoren beeinflußt waren:

Schüler waren unmotiviert für die freiwillige Testdurchführung und erzielten deshalb schlechtere Ergebnisse, als sie sie aufgrund ihrer tatsächlichen Kompetenz erbringen können.

Schüler waren nicht bereit, Leistungen selbständig zu erbringen und erzielten durch Abschreiben bessere Leistungen, als sie ihrer tatsächlichen Kompetenz entsprachen.

Gelegentlich wurde auch beobachtet, daß die CT-Ergebnisse besser als die Schulnote waren, da der CT nicht durch ein negatives interpersonelles Verhältnis zwischen Lehrer und Schüler beeinträchtigt wurde, wie das bei der Schulnote leider der Fall zu sein schien.

Diese Erkenntnisse führten zu der Einsicht, daß die Schulzensur nicht das einzige Validierungskriterium sein durfte, sondern ergänzt werden müsse durch eine Lehrerbeurteilung, die alle für die Fremdsprachenkompetenz irrelevanten Faktoren auszuschalten habe.
Schwierigkeiten bereitete die Interpretation der ersten Reliabilitäts- und Validitätskoeffizienten. Es waren keine klaren Strukturen erkennbar. Es wurde argumentiert, daß die Aussage eines Einzeltests pro Klasse hierzu nicht ausreiche, zumal der Test unter Umständen zu neuartig war und damit Fehlerquellen, bedingt durch die Unbekanntheit der Testmethode, die ja ohne Zweifel auch für die Schüler einen gewissen Neuigkeitscharakter beinhaltete, entstehen konnten.
Es wurde auch die Möglichkeit in Betracht gezogen und erwogen, daß die Testlänge nicht ausreiche, um zuverlässige Ergebnisse zu erzielen.
Für die Hauptuntersuchung wurde deshalb - diese Bedenken und Überlegungen berücksichtigend - die Durchführung von drei Tests verschiedener Schwierigkeitsgrade in jeder Klasse beschlossen, um zuverlässige Aussagen zu erhalten und Leistungsstrukturen explizit zu machen.

Nicht alle Tests wurden von den Schülern gleich positiv aufgenommen. Einige wurden wegen der Textinhalte weniger wohlwollend und deshalb kritisiert. Dazu gehörten z.B. ein Text, der wegen seines technischen Titels, "How Television Works", vor allem bei Mädchen auf Ablehnung stieß, und ein politisch orientierter Text, der ebenfalls von derselben Schülergruppe abgelehnt wurde. Ein anderer Test wurde wegen seines unnatürlichen, lehrbuchmäßigen Textes mit Widerwillen bearbeitet. Diese Texte wurden infolgedessen ausgesondert.

Eine Reihe von anderen Variablen wurde in diesen Vorversuchen eruiert. Während in der Literatur die meisten Tests 50 Aufgaben enthielten, wurden 25 Aufgaben für günstig befunden, da der Test dann von Klassen mit niedrigerem Kompetenzniveau ohne Hetze innerhalb einer Schulstunde bearbeitet werden konnte, und bei besseren Klassen der Test während einer Stunde noch durchbesprochen werden konnte, um ein unmittelbares Feedback - ein als sehr günstiges in der Lernpsychologie bezeichnetes Moment - zu erzielen.
Ein Test dieser Länge paßte auch genau auf ein DIN A 4 - Blatt und ließ noch genügend Platz zur Datenerfassung und Korrektur (siehe auch Beispiel 1 und 2 im Kapitel 2.4.1. von Hinofotis und Brown).
Dieses Format erwies sich als praktikabel und sollte für die Hauptversuche

beibehalten werden.

Das klarste Ergebnis wurde bei den Vorversuchen in Bezug auf die A u s w e r t u n g s m e t h o d e festgestellt. Da die 'exact-scoring'-Methode von verschiedenen Autoren (z.B. Stubbs & Tucker, 1974) empfohlen wurde, wurde sie zunächst in der Erprobung verwendet. Sie stieß jedoch bei den Probanden auf Ablehnung, da es für sie nicht einsichtig war, warum aus rein statistischen Gründen nur die Wörter aus dem Originaltext, nicht aber jene, die auch sinngemäß richtig waren, gezählt werden sollten. Warum sollte z.B. ein Proband, der im Testbeispiel von Brown (siehe Kapitel 2.4.1.)

"There is no race of men _(1)_ on earth so backward that it _(2)_ no language, ..."

anstatt "anywhere" für (1) und "has" für (2) nicht auch "known" für (1) und "knows" für (2) als richtig einsetzen können? Wäre dies nicht möglich würde die psychologische Gültigkeit und die mit ihr verbundene Probandenmotivation stark beeinträchtigt und in Mitleidenschft gezogen.

Um dem Probanden größere Gerechtigkeit widerfahren zu lassen, wurde eine Auswertungsmethode erprobt, die auch die Schwere der Fehler berücksichtigte. Die Erwägung ging soweit, daß je nach der Schwere des Fehlers die Antwort mit unterschiedlichen Stellenwerten bedacht wurde und je nach dem mit 4, 3, 2 oder 1 bzw. 0 Punkten gewichtet wurde.[24]
Die Fehler wurden wie folgt gewertet:

4 PUNKTE: semantisch und grammatisch akzeptable Antwort
3 PUNKTE: semantisch richtige, aber grammatisch falsche Antwort;
2 PUNKTE: semantisch falsche, aber grammatisch richtige Antwort;
1 PUNKT: richtige Wortklasse.

Es ergab sich eine Reihe von Schwierigkeiten. In vielen Fällen war es zunächst nicht absehbar, welche Wortklasse akzeptabel war, bzw. den Begriff der "Wortklasse" in akzeptabler Weise zu definieren.
Ein Beispiel aus dem Test "Who is Who?" zeigt, daß für die Aufgabe 13 akzeptable Antworten mehreren Wortklassen angehören können:

"You see, Ruth, Robert says, that's _(13)_ I say, her name is Barbara."

Im Hinblick auf den gesamten restlichen Kontext sind hier folgende Antworten mit Wörtern verschiedener Wortklassen möglich:

what / right / Barbara

Möglicherweise gibt es noch weitere akzeptable Antworten mit noch anderen

---

24 Diese Anregung verdanke ich Herrn Prof. Hecht.

Vgl. auch F.W. Radice in: ELT, No.3, 1978, der sogar eine siebenstufige Fehlerbewertung - allerdings nicht in Tests sondern in Übungen - erprobt hat,

Wortklassen. Dies ist übrigens kein Einzelbeispiel; im gleichen Text sind noch zwei weitere Beispiele ähnlicher Art (17 und 20).

Weiß der Korrektor nicht genau im voraus, welche Wortklassen als Antwort akzeptabel sind, und wie er sie definiert, bzw. die Zugehörigkeit einzelner Wörter zu den Wortklassen bestimmt, ist hier kaum eine saubere Beurteilungsweise möglich. Der Beurteiler müßte in diesem Fall praktisch für jedes Adjektiv, wenn man das im Sinne von Fries (1952) als die Wortklasse von 'right' anerkennen will, einen Punkt geben. Da aber im Englischen die meisten Adjektiva zusätzlich grammatisch richtig wären, müßte er sogar zwei Punkte erteilen.
Vergleicht man als Antworten z.B. "funny" und "black", könnte man intuitiv behaupten, "funny" ist richtiger als "black", bzw. "black" ist völlig abwegig und sollte nicht einmal einen Punkt erhalten, obwohl es konsequenterweise nach der oben erwogenen Regelung mit zwei Punkten einzustufen wäre. In der Konsequenz könnte sich ein negativer 'Backwasheffekt' einstellen. Kluge Schüler würden im Zweifelsfall nur noch ein semantisch wie immer auch geartetes Wort einer grammatisch intuitiv möglichen Form einsetzen, um zwei Punkte dafür zu erhalten. Es ist fraglich, ob dadurch tatsächlich die fremdsprachliche Kompetenz überprüft wird oder ob es nur eine raffinierte Ratetechnik wäre.
Die Fortführung dieser linguistischen Diskussion, die aber hier nicht weiter relevant ist, könnte das Thema einer eigenen Untersuchung sein. In diesem Zusammenhang ist zu schließen, daß diese "wertende Methode" auch insofern unpraktikabel ist, als sie zusätzlich den technischen Nachteil mit sich bringt, daß die Verwendung eines Korrekturschlüssels zur raschen Auswertung so gut wie unmöglich ist.

Als beste Möglichkeit erschien die 'acceptable word'- Methode, wie das auch vom statistischen Standpunkt her gesehen aus der Literatur empfohlen wird.

Als undurchführbar erwies sich der Versuch, die von den Schülern im Unterricht geschriebenen Testarbeiten als Vergleichskriterien heranzuziehen. Die Idee war zunächst, daß - ähnlich den Subtestkorrelationen der vorgestellten amerikanischen Studien - besondere Korrelationen zwischen dem CT und Testarbeiten der Schüler, wie etwa Diktat und Leseverstehen, aufgedeckt werden könnten. Die wenigsten der Klassentests konnten eindeutige Aussagen über besondere Teilkompetenzen machen, bzw. sie konnten

nicht in der Richtung von Teilkompetenzen festgelegt werden. Das bedeutet allerdings nicht, daß Klassentests immer oder vorwiegend integrative Tests im hier verwendeten Sinne sind.
Das Gegeteil dürfte die Regel sein.
In den meisten Fällen setzten sich Klassenarbeiten aus Teilarbeiten etwa zur Grammatik, zum Vokabular, zum Leseverstehen, Diktat etc. zusammen; es war aber hier nicht möglich, diese Detailinformationen separat zu erhalten und auszuwerten. Auch hier zeigt sich die Notwendigkeit zukünftiger Erforschung der Qualität von Klassentests.

Als weiterer Problempunkt erwiesen sich die Sprachfähigkeiten der Schüler aus Gastarbeiterfamilien. Das Bild ist hierbei sehr vielschichtig. Ausländische Kinder an Gymnasien und Realschulen leisten häufig im Englischunterricht mehr als ihre deutschen Mitschüler. Es kann sein, daß sie besondere Sprachlernstrategien entwickelt haben, da Englisch nicht ihre erste Fremdsprache, sondern ihre zweite, manchmal auch dritte Fremdsprache ist. Auch mitunter an Hauptschulen war dies beobachtbar. Es war aber auch der umgekehrte Fall zu sehen: Schüler waren so gut wie ausgeschlossen von Fremdsprachenunterricht, weil sie die Doppelgleisigkeit des Englischunterrichts, der ja auch zu einem wichtigen Teil in Deutsch erteilt wird, nicht bewältigen konnten. Probanden, bei denen Englisch die Drittsprache ist, wurden wegen der besonderen Problematik von der Untersuchung ausgeschlossen. Zukünftige Forschung scheint hier nötig zu sein.

Bei den Vorversuchen wurde festgestellt, daß keiner der insgesamt 16 entworfenen Tests von allen Gruppen komplikationslos bearbeitet werden konnte. Leistungsfähige Schülergruppen waren weit unterfordert von Tests, die für Gruppen mit einem niedrigeren Kompetenzniveau gerade angemessen waren, und diese wiederum waren weit überfordert von angemessenen Tests für fortgeschrittenere Gruppen. Um zu vermeiden, daß Tests nicht von allen Probanden bearbeitet werden konnten und somit keine Vergleichbarkeit von Ergebnissen mehr möglich war, schien es für das Hauptexperiment vorteilhaft zu sein, sich auf eine homogenere Probandengruppe zu beschränken. Ein zeitweise ins Auge gefaßter Plan, nur die 9. Jahrgangsstufe zu untersuchen, wurde fallen gelassen, da die Ergebnisse zu wenige Vergleichbarkeiten und Vergleichsmöglichkeiten boten, und die Aussagen einen zu spezifischen Charakter hätten annehmen müssen. Als günstigste Möglichkeit wurde für das Hauptexperiment deshalb die

Untersuchung der Jahrgangsstufen 5 bis 9 beschlossen.

4.2. Forschungshypothesen

Die allgemeine Forschungshypothese dieser Untersuchung ist, daß das Standardverfahren des CT ein brauchbares Meßinstrument der allgemeinen Fremdsprachenkompetenz bestimmter deutscher Schülergruppen ist, was empirisch nachgewiesen werden soll durch die Gütekriterien der Reliabilität, Validität und Praktikabilität.
Folgende Einzelhypothesen wurden aufgestellt:

HYPOTHESE I: Der CT ist in der Lage schul- und jahrgangsspezifische Fremdsprachenkompetenzunterschiede aufzuzeigen.

HYPOTHESE II: Der CT ist ein reliables Meßinstrument, ausgedrückt durch akzeptable Reliabilitätskoeffizienten.

HYPOTHESE III: Der CT ist ein gültiges Meßinstrument der Fremdsprachenkompetenz, ausgedrückt durch akzeptable Korrelationskoeffizienten mit der Beurteilung der fremdsprachlichen Kompetenz durch den Lehrer.

HYPOTHESE IV: Der CT ist ein praktikables Meßinstrument der Fremdsprachenkompetenz, ausgedrückt durch

a) den Arbeits- und Zeitaufwand bei der Erstellung des Tests;
b) den Arbeits- und Zeitaufwand bei der Testdurchführung;
c) den Arbeits- und Zeitaufwand bei der Auswertung des Tests.

4.3. Plan zur Durchführung der Studie

Folgender Plan wurde zur Überprüfung der aufgestellten Hypothesen gefaßt:
Drei Tests verschiedener Schwierigkeitsgrade werden in verschiedenen Hauptschul-, Realschul- und Gymnasialklassen durchgeführt. Gleichzeitig werden die Schulnoten und die Lehrerbeurteilung der Fremdsprachenkompetenz der Schüler erfaßt.
Die Ergebnisse der Testdurchführung, d.h. die arithmetischen Mittelwerte der Gruppen und Standardabweichungen, werden zur Überprüfung der Hypothese I zur schulspezifischen Lage der Fremdsprachenkompetenz herangezogen. Diese Ergebnisse können aber auch zur Beantwortung der Fragen nach Aus-

wirkungen der Textschwierigkeit herangezogen werden, und sie können Aussagen zur Gültigkeit der Tests liefern.
Die Überprüfung der HYPOTHESE II geschieht durch die Berechnung der Reliabilitätskoeffizienten. Nach den aus der Literaturübersicht gewonnenen Erkenntnissen und der danach getroffenen Entscheidung werden die Split-half-Methode mit der Spearman-Brown-Formel für r_{tt} und die Berechnung nach Kristof verwendet.
HYPOTHESE III wird durch Berechnung der kriterienbezogenen Validität überprüft. Sie ergibt sich aus der Korrelation zwischen den CT-Ergebnissen einerseits und den beiden Kriterien Englischnote und Lehrerbeurteilung andererseits. Das angewandte Verfahren ist die PEARSON-PRODUKT-MOMENT-KORRELATION.
HYPOTHESE IV zur Praktikabilität wird direkt durch die Beobachtungen bei der Testdurchführung überprüft.

Aus diesen Ergebnissen sollen Schlüsse zur Beantwortung der eingangs gestellten Forschungsfragen gezogen werden.

4.4. Die Probanden

Die Ergebnisse der Hauptuntersuchung stammen von N = 440 Schülern.
Davon sind 148 Schüler von drei verschiedenen Hauptschulen, 61 Schüler einer Realschule und 266 Schüler von zwei verschiedenen Gymnasien.
Sie stammen aus 10 Hauptschulklassen, 8 Gymnasialklassen und 3 Realschulklassen.
Von den Hauptschulklassen war eine Klasse gemischt (A und B Kurs), sechs Klassen waren A Kurse, drei Klassen B Kurse. (Die Kurszugehörigkeit wird durch die Zensuren bestimmt). Die jeweils besseren Schüler waren in den A Kursen, die schlechteren in den B Kursen.
Eine detaillierte Aufschlüsselung der Probandenzahl ist in der TAFEL 19 ersichtlich.
Die ursprüngliche Gesamtzahl der Schüler und Klassen war beträchtlich höher. Da für die Untersuchung aber die Durchführung von drei Tests bei jedem Schüler und in jeder Klasse vorgesehen war, gab es beträchtliche Ausfälle. Jeder Schüler, der auch nur bei einem Test fehlte, reduzierte die Gesamt-

TAFEL 19

Schulart / Jahrgang	Schule / Klasse	N
HS 5	LO 5a	14
GY 5	BB 5c	29
GY 5	OT 5b	26
HS 6	FFB 6e A Kurs	16
GY 6	BB 6b	26
GY 6	OT 6e	32
HS 7	FFB 7d A Kurs	10
HS 7	HA 7abc B Kurs	19
RS 7	A 7a	28
GY 7	OT 7e	30
HS 8	FFB 8d A Kurs	12
HS 8	HA 8ab A Kurs	18
HS 8	HA 8ab B Kurs	18
RS [illegible]	[illegible]	[illegible]
GY 8	OT 8f	28
HS 9	LO 9ab A Kurs	21
HS 9	LO 9ab B Kurs	10
HS 9	FFB 9c A Kurs	10
RS 9	A 9f	18
GY 9	OT 9d	31
GY 9	OT 9e	29

Erklärung:

Schulart / Schülerjahrgang:

HS 5 = Hauptschule, 5. Jahrgang

GY 5 = Gymnasium, 5. Jahrgang

RS 7 = Realschule, 7. Jahrgang

Schule / Klasse:

LO 9ab = Ort, Klassen- und
A Kurs Kursbezeichnung

N = Anzahl der Probanden

zahl der zur Verfügung stehenden Probanden seiner Klasse. Teilweise fielen auch Klassen ganz aus, die nicht für alle drei Tests zur Verfügung stehen konnten. Da solche Ausfälle unvorhersehbar waren, konnte nicht immer die wünschenswerte Stichprobengröße erreicht werden.
Am häufigsten waren Ausfälle an den Hauptschulen zu beklagen; dort kam des öfteren auch vor, daß Klassen nach den ersten zwei Tests überhaupt nicht mehr zur Verfügung standen, weil mitunter bereits ab Mai / Juni kein regulärer Englischunterricht mehr stattfand.
Diese Tatsache ist indikativ für den Stellenwert des Englischunterrichts an den verschiedenen Schularten anzusehen.
Die Auswahl der einzelnen Schulen und Klassen war insofern vom Zufall bestimmt, als die jeweiligen Englischlehrer dem Autor bekannt waren, und die Klassen für die Untersuchung zur Verfügung stellten.
Die Schulen befinden sich alle in der näheren und weiteren Umgebung von München und können als repräsentativ für die jeweilige Schulart gesehen werden, soweit das möglich ist.
Bei den Probanden selbst wurden keine außergewöhnlichen Tatsachen beobachtet. Aus den oben dargestellten Gründen nahmen die Schüler mit anderen Muttersprachen als Deutsch am Test nicht teil. Die Teilnahme war für alle Schüler freiwillig.

4.5. Die Testkonstruktion

Die Testkonstruktion des CT erfolgte in vier Schritten:

1) Eine in Stil, Schwierigkeit, Inhalt und Länge angemessene Textpassage wurde ausgewählt.

2) Lücken im Abstand von 7 Wörtern wurden gebildet.

3) Das Testblatt mit den Anweisungen für die Probanden und den Räumen für die Antworten wurde zusammengestellt.

4) Der vierte Schritt bestand in der Ausarbeitung der Antwortschablonen.

Die TEXTAUSWAHL: Sie erforderte einen relativ großen Aufwand, um alle Bedingungen zu erfüllen, die an den Text gestellt wurden. Er sollte von Stil und Inhalt her zwar neutral aber doch auch anregend für die Schüler sein. Der Text sollte innerhalb von ca 180 Wörter eine abgeschlossene Einheit bilden und einen angemessenen Schwierigkeitsgrad für die Testpopulation haben.
Am geeignetsten erwiesen sich Lehrbuchtexte, von denen jene ausgesondert werden mußten, deren Stil zu "lehrbuchmäßig" bzw. unnatürlich war. Nach den Ergebnissen der Vortests wurden folgende drei Texte als geeignet befunden:

1) "Who is Who" aus "English for You" I von Knippschild (1975:48)
   Nach der sog Fry-Readability-Chart (Fry, 1977) hat dieser Text einen Schwierigkeitsgrad zwischen '1st and 2nd grade'.

2) "Pat at the Grocer's" aus "The Good Companion" von Rohr, Heinz et al (^{2}1970:28) mit einem Schwierigkeitsgrad nach Fry von 2nd grade.

3) "Tim's First Drink" aus "English H5" von Friedrichs, Hildegard (^{2}1975:48) mit einem Schwierigkeitsindex nach Fry zwischen 2nd und 3rd grade.

Alle drei Texte enthalten sowohl dialogische als auch narrative Elemente, womit (nach Wilkins, ^{2}1977:23ff) die beiden Sprachfunktionen des 'Reporting & Doing' abgedeckt sind.

Der LÜCKENABSTAND:
Aus diesen Texten wurde nach einem Einleitungssatz zur Orientierung des Schülers (vgl. Oller, 1979:365) jedes 7. Wort ausgelassen.

Das Verfahren wurde dabei nicht ganz präzise eingehalten, denn einige offene Fragen behinderten die strikte Durchführung. Folgende Fragen wurden gestellt:

"Wie sind Eigennamen, z.B. 'The Ship of Gold' zu zählen?"

"Wie sind Kontraktionen, z.B. 'I'm' zu zählen?"

"Ist ein Ausruf, z.B. 'Oh' ein Wort?"

Im Zusammenhang dieser Tests schien die Beantwortung dieser technischen Fragen jedoch unbedeutend, solange eine zufällige Lückenauswahl in einem in etwa konstanten Abstand getroffen wurde.

Die BEREITSTELLUNG DER ANTWORTRÄUME:

Außer dem Testtext sollte das Blatt noch Platz bieten für die Schülerantworten, die Testanweisungen und einen Raum zur Erfassung von Information über die Probanden. In den Vorversuchen hat sich gezeigt, daß es am günstigsten und leichtesten für den Schüler war, gleich unmittelbar auf das Testblatt zu schreiben und nicht auf ein separates Antwortblatt, wie das sonst häufig üblich ist.

Die TESTANWEISUNGEN:

Die Anleitungen sollten leicht verständlich und eindeutig sein. Bei den Vorversuchen hatte sich die schließlich verwendete Form als die beste gezeigt. Die Schüler wurden aufgefordert, den Text erst ganz durchzulesen und dann die Wörter, die sie für passend hielten, einzusetzen. Nachdrücklich wurde auf die "Spielregeln" des CT hingewiesen, nur ein Wort einzusetzen, und daß die Rechtschreibung unberücksichtigt bliebe, solange das Wort für den Auswerter erkennbar wäre. Damit sollte vermieden werden, daß orthographische Fertigkeiten die Aussage zur Fremdsprachenkompetenz beeinflussen.

Die Informationen über den Probanden sollten gegebenenfalls auch für zukünftige Fragestellungen Auskunft geben können, etwa zu der Frage, ob die Deutschnote mit der Englischnote korreliert im Hinblick auf eine allgemeine Sprachkompetenz.

Was die Angabe des Probandennamens auf dem Testbogen betrifft, so diente dies nur dazu, die drei verschiedenen Tests im Hinblick auf das Lehrerurteil einander zuordnen zu können. Die Schüler wurden darauf hingewiesen, daß ihre Namensangabe freiwillig sei und alle Namen nach Beendigung der Untersuchung beseitigt würden.

Die AUSARBEITUNG DER ANTWORTSCHABLONEN:

Aus den Vortests lag eine Reihe möglicher Antworten zu den einzelnen Aufgaben vor. Wo Zweifel an der Akzeptabilität der Antworten bestanden, wurden

sie einer Reihe von native speakers vorgelegt.[24]
In allen Fällen wurde Übereinstimmung zur Annahme oder Verweigerung zweifelhafter Antworten erzielt; sie wurden auf dem Lösungsblatt so angeordnet, daß der Schlüssel nur neben das Arbeitsblatt gelegt werden brauchte.

4.6. Die Erhebung der Validierungskriterien

Nachdem sich in den Vorversuchen herausgestellt hatte, daß in die Schulnoten unter Umständen Faktoren einfließen können, die für eine Aussage zur Fremdsprachenkompetenz irrelevant sind, wurde neben den Schulnoten, von denen jeweils die letzte herangezogen wurde, auch das Lehrerurteil zur Fremdsprachenkompetenz der Schüler eingeholt.
Hierzu wurden die Lehrer gebeten, so neutral wie möglich, ihre Urteile zur allgemeinen Fremdsprachenkompetenz der Schüler in der üblichen Form der sechsstufigen Zensur anzugeben. Insbesondere sollten sie alle Einflüsse, die nicht relevant zur tasächlichen Fremdsprachenkompetenz der Schüler sind, die aber die Zeugniszensuren beeinflußt haben könnten, ausklammern.

4.7. Die Testdurchführung

Die Versuche wurden in den Schuljahren 1978/79 bis 1980/81 durchgeführt. Die Zeitpunkte der Durchführung (Z) und die Sequenz der einzelnen CTs sind aufgeführt in TAFEL 20. Die Abstände zwischen den in einer Klasse durchgeführten Einzeltests betrugen zwischen ein paar Tagen und einigen Wochen, gelegentlich auch einigen Monaten. Die Unregelmäßigkeit dieser Abstände beruhte darauf, daß die Klassen nicht regelmäßig zur Verfügung standen, sondern immer nur dann, wenn der reguläre Englischunterricht nicht gestört wurde. Aus diesen Gründen mußten auch ungünstige Zeiten wie z.B. die letzte Stunde eines Unterrichtstages in Kauf genommen werden.

---

24 An dieser Stelle bin ich Mr Mc Credie, Miss Evans, Dr. Margulis, Mr Look, Mr Marks, Mr Dera, Miss Stevee und Mr Davies zu Dank für ihre Hilfe verpflichtet

TAFEL 20

DURCHFÜHRUNGSDATEN VON 3 CTs IN DEN JAHRG.ST. 5-9 VERSCHIEDENER HS, RS, GY

Test / Schulart Jahrgang	Schule Klasse	N	I Ts	I Z	II Ts	II Z	III Ts	III Z
HS 5	LO 5a	14	1	7/80	2	7/80	3	7/80
GY 5	BB 5c	29	1	3/79	3	7/79	2	7/79
GY 5	OT 5b	26	1	9/79	2	9/79	3	9/79
HS 6	FFB 6e A Kurs	16	1	5/79	2	10/79	3	10/79
GY 6	BB 6b	26	1	3/79	2	3/79	3	9/79
GY 6	OT 6e	32	1	7/79	2	7/79	3	7/79
HS 7	FFB 7d A Kurs	10	1	7/79	2	7/79	3	12/79
HS 7	HA 7abc B Kurs	19	1	5/79	2	5/79	3	7/79
RS 7	A 7a	28	2	4/79	3	7/79	1	9/79
GY 7	OT 7e	30	1	7/79	2	7/79	3	7/79
HS 8	FFB 8d A Kurs	12	1	2/79	2	2/79	3	7/79
HS 8	HA 8ab A Kurs	18	1	7/79	2	7/79	3	7/79
HS 8	HA 8ab B Kurs	18	1	7/79	2	7/79	3	7/79
RS 8	A 8a	15	2	7/79	3	9/79	1	1/79
GY 8	OT 8f	28	3	9/79	2	9/79	1	4/79
HS 9	LO 9ab A Kurs	21	1	12/79	2	12/79	3	12/79
HS 9	LO 9ab B Kurs	10	1	10/79	2	12/79	3	2/80
HS 9	FFB 9c A Kurs	10	1	10/79	2	1/80	3	1/80
RS 9	A 9f	18	1	9/80	2	9/80	3	9/80
GY 9	OT 9d	31	1	7/80	2	7/80	3	7/80
GY 9	OT 9e	29	1	7/80	2	7/80	3	7/80

Vor der Durchführung erhielten die Probanden alle die gleiche Information:

1) Der Test soll die fremdsprachlichen Fähigkeiten der Schüler messen.
2) Der Test dient der wissenschaftlichen Forschung.
3) Die Teilnahme am Test ist freiwillig.
4) Die Arbeiten werden nicht benotet.
5) Die bestmögliche Leistung soll erreicht werden.
6) Jeder Schüler soll selbständig arbeiten.
7) Die Ergebnisse bleiben anonym.
8) Jeder Schüler hat genügend Zeit zur Lösung.

Danach wurde die Testanleitung, die sich auf dem Testbogen befindet, gemeinsam besprochen und Fragen, die auftauchten, dazu beantwortet.
Mit der Testbearbeitung waren die Schüler je nach Klassenstufe und Schulart in 5 bis 25 Mintuen fertig. Danach wurden die Antworten besprochen. Dies erwies sich als sehr nützlich, denn hier konnte sofort Klärung geschaffen werden, wenn irgendwelche Probleme aufgetaucht waren. Insbesondere konnte sofort eine Über- oder Unterforderung festgestellt werden. Der Lehrer erhielt auch detaillierte Einblicke in die Fähigkeiten seiner Schüler . Von diesem Standpunkt ist dem CT ein gewisser diagnostischer Wert nicht abzusprechen.
Es konnten sehr gut feine Nuancen in den Unterschieden zwischen verschiedenen Antworten erarbeitet bzw. besprochen werden. Die Schüler waren an diesen Besprechungen rege beteiligt, denn durch ihre vorherige Beschäftigung und Auseinandersetzung mit dem Testinhalt waren sie stark motiviert.
Für die didaktische Verwertung des CT sei dieser Teil als wichtiger und ergiebiger Punkt empfohlen.

Trotz der Aufforderung zur selbständigen Arbeit und der Versicherung, daß die Testergebnisse anonym bleiben würden, wurde festgestellt, daß Schüler von ihrem Nachbarn abschrieben. Da das Format des Tests mit den deutlich sichtbaren und geordneten Antworten die Schüler zum Abschreiben geradezu reizte, erweist sich eine besonders sorgfältige Aufsicht als angeraten, um eine Verfälschung der Ergebnisse zu vermeiden.

Die Tests wurden mit Hilfe der Antwortschablonen ausgewertet. Die Auswertung bzw. Korrektur nahm weniger als eine Minute pro Arbeitsbaltt in Anspruch. Die Rohwertergebnisse der einzelnen Schüler und Klassen sind aufgeführt im Appendix.

Die Berechnungen der statistischen Werte erfolgten mit Hilfe eines Kleincomputers (TI - 51 - III).

5. RESULTATE

In diesem Kapitel werden die statistischen Daten präsentiert und erklärt, die aus den bei der Testdurchführung erhaltenen Rohwerten berechnet wurden.

5.1. Mittelwerte, Schwierigkeitsindizes und Streuungswerte

Die aus den Rohwerten der CTs berechneten arithmetischen Mittel ($\bar{X}_{ct}$) zeigen die jeweils erreichten Leistungen im Klassendurchschnitt. Die Mittelwerte wurden für jeden der drei Tests in allen Jahrgängen und Schularten errechnet. Daneben wurden für jede Klasse auch die arithmetischen Mittel der Testbatterie aus den drei CTs berechnet ($\bar{X}_{ct\ 123}$). Zur Berechnung wurde die Formel 1 verwendet.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

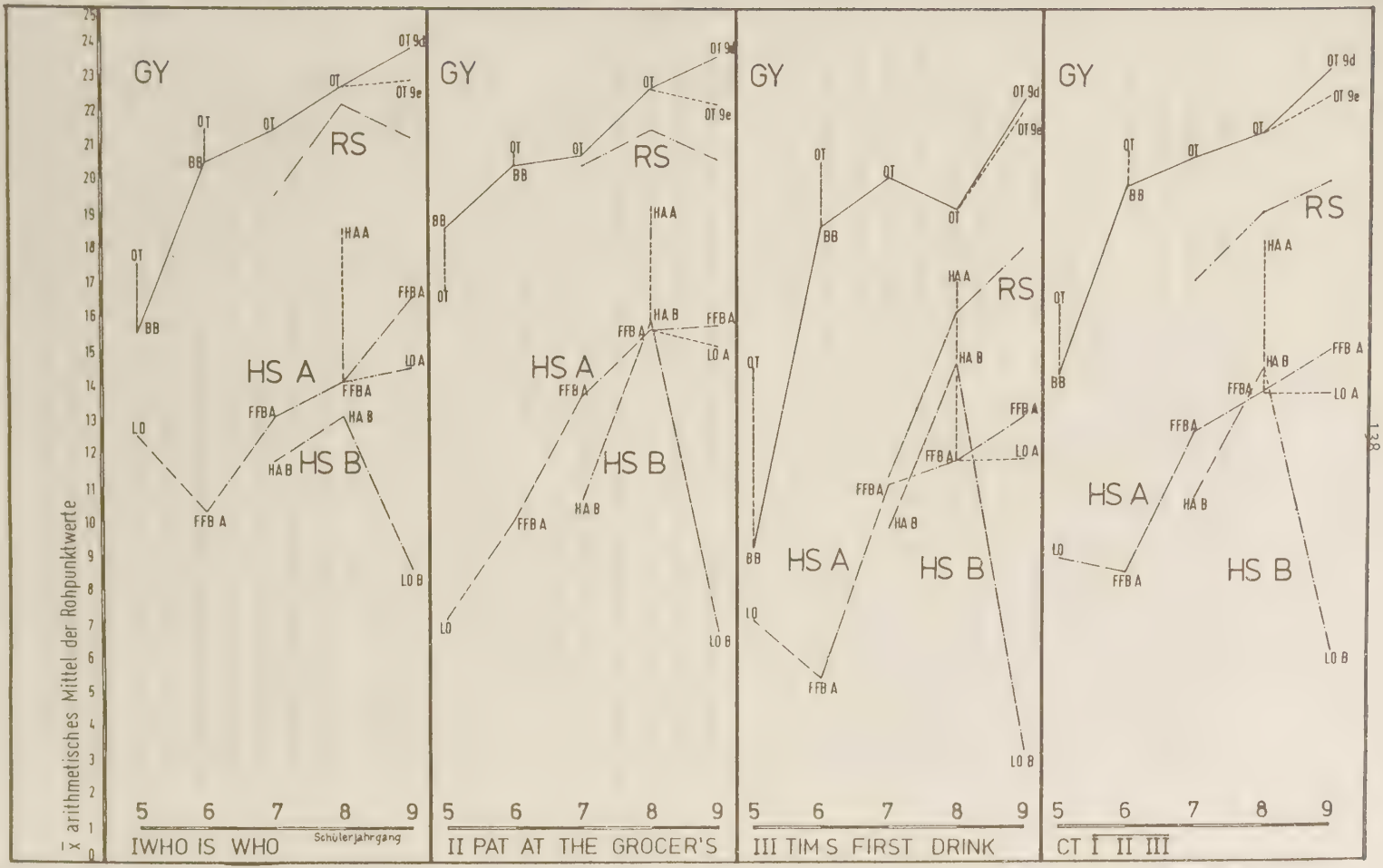
Dabei ist n die Anzahl der Meßwerte.
TAFEL 21 zeigt die arithmetischen Mittelwerte.
TAFEL 22 veranschaulicht die arithmetischen Mittelwerte und ist somit eine graphische Repräsentation dieser Ergebnisse.

TAFEL 21

Arithmetische Mittel der Einzeltests und der Gesamttestbatterie in allen ~~Klassen~~ Schularten

Test / Schulart / Jahrgang	Schule Klasse	N	I $\bar{X}_{CT}$ RANGE	II $\bar{X}_{CT}$ RANGE	III $\bar{X}_{CT}$ RANGE	CT I+II+III $\bar{X}_{CT}$ RANGE
HS 5	LO 5a	14	12,5 7-17	7,07	7,4	8,9 4,7
GY 5	BB 5c	29	15,48 10-21	[illegible]	[illegible]	[illegible]
GY 5	OT 5b	26	17,62 7-24	[illegible]	[illegible]	[illegible]
HS 6	FFB 6e A Kurs	16	10,313 6-18	[illegible]	[illegible]	[illegible]
GY 6	BB 6b	26	20,54 16-24	[illegible]	[illegible]	[illegible]
GY 6	OT 6e	32	21,50 14-25	[illegible]	[illegible]	[illegible]
HS 7	FFB 7d A Kurs	10	13,10 9-17	[illegible]	[illegible]	2,09 [illegible]
HS 7	HA 7abc B Kurs	19	11,84 4-22	[illegible]	[illegible]	[illegible]
RS 7	A 7a	28	19,46 13-23	[illegible]	[illegible]	[illegible]
GY 7	OT 7e	30	21,37 8-25	[illegible]	[illegible]	[illegible]
HS 8	FFB 8d A Kurs	12	14,7 10-19	[illegible]	[illegible]	13,858 [illegible]
HS 8	HA 8ab A Kurs	18	18,07 15-22	[illegible]	[illegible]	18,272 [illegible]
HS 8	HA 8ab B Kurs	18	[illegible]	[illegible]	[illegible]	14,578 [illegible]
RS 8	A 8a	17	[illegible]	[illegible]	[illegible]	[illegible]
GY 8	OT 8f	28	22,82 18-25	[illegible]	[illegible]	21,268 17,6-24,3
HS 9	LO 9ab A Kurs	21	14,48 4-23	[illegible]	11,76 1-20	13,738 3,3-20,6
HS 9	LO 9ab B Kurs	10	8,6 3-16	6,7 1-12	3,2 0-6	6,11 4,3-9,3
HS 9	FFB 9c A Kurs	10	16,6 8-22	[illegible]	13,1 8-20	15,05 9,6-20,6
RS 9	A 9f	18	21,17 12-25	20,56 18-24	18,06 8-22	19,93 13-22,6
GY 9	OT 9d	31	23,78 19-25	[illegible]	22,32 19-25	25,27 19,4-29,7
GY 9	OT 9e	29	22,97 18-25	[illegible]	22,21 18-25	22,45 20,3-24,7

LEISTUNGSPROFILE



Diese Ergebnisse entsprechen den Aussagen, die in Kapitel 2.3. über die schulischen Leistungsstrukturen gemacht wurden. Es werden deutlich die Leistungsunterschiede zwischen den Schularten gezeigt. Bei allen drei Tests sowie bei der Kombination der drei Tests (CT I II III) sind konstant die Mittelwerte der Gymnasialklassen höher als die der Realschulklassen und diese wiederum höher als die der Hauptschulklassen.
Hierbei wiederum sind die A-Kurswerte höher als die B-Kurswerte. Das Muster wird jedoch bei CT III und bei CT I II III durchbrochen: der Hauptschul - A - Kurs HA 8ab ist besser als die Realschulklasse RS 8a. Auch der Hauptschul - B - Kurs HA 8ab ist wesentlich besser als die HS FFB 8d A - Kurs - Vergleichsklasse. Der Grund lag darin, wie sich nach der Testdurchführung herausstellte, daß den beiden Gruppen HA 8ab A- Kurs und B - Kurs der Testtext (Tim' s First Drink) vom Unterricht her bekannt war.[24a] Konsequenterweise macht sich der Einfluß dieses überdurchschnittlichen Mittelwertes auch bei CT I II III bemerkbar, wo der B - Kurs HA 8ab besser ist als der A- Kurs FFB 8d.
Es sei angemerkt, daß die Klasse 8ab eine sehr stark geförderte und motivierte Klasse war, die u.a. bereits zur Unterrichtsdemonstration im Fernsehen herangezogen wurde. Ihr überdurchschnittlicher Leistungsstand wird deutlich durch den CT erfaßt.
Kleinere Abweichungen vom Muster sind leicht durch normale Leistungsschwankungen erklärbar. Es kann auch sein, daß ein Test einer Klasse besonders leicht oder besonders schwer fiel, weil ihr bestimmte Voraussetzungen fehlten. Dies war z.B. der Fall bei der Klasse HS LO 5a in Test II.
Die Klsse war in den Tests I und II relativ besser als die 6. Klasse. Schwierigkeiten mit dem Vokabular führten dazu, daß hier das Muster durchbrochen wurde.
Zwischen den vier Graphiken der TAFEL Nr. 22 ist deutlich eine Ähnlichkeit der Verteilungsmuster der Mittelwerte erkennbar.

Die Muster selbst entsprechen ebenfalls den Darstellungen der Leistungsstrukturen aus Kapitel 2.3.. Die Leistungen steigen an, je höher die Schülerjahrgänge sind. Diese Tendenz stimmt generell für alle drei Schularten, wird aber aus bestimmten Gründen an einigen Stellen unterbrochen. Die Hauptschulklasse LO 5a erzielte höhere Werte als die 6. Klasse (mit der oben gezeigten Ausnahme). Im Rückgriff auf die Darstellung der Zuordnung der Schüler an die Schularten sei wiederholt, daß im 5. Schüler-

---

24a Dies war unvermeidbar, da ein Lehrerwechsel stattgefunden hatte und die Schüler selbst offensichtlich ihren Vorteil nicht preisgeben wollten.

jahrgang noch ein Teil zukünftiger leistungsfähiger Gymnasiasten in der Hauptschule bleibt und erst nach dem 5. Schuljahr an das Gymnasium übertritt. Das erklärt, warum die 5. HS-Klasse bessere Leistungen erzielte als die 6. HS-Klasse.
Der B - Kurs der Klasse LO 9ab erzielte die schlechtesten Leistungen insgesamt. Ein wesentlicher Faktor dürfte mangelnde Motivation sowohl auf der Seite der Schüler als auch auf der Seite des Unterrichtenden sein. Es sei hier angmerkt, daß ein Teil der Testbogen verschmiert und zerknittert zurückgegeben bzw. im Klassenzimmer gefunden wurde. Die Klasse hat auch mit 4,6 (!) den schlechtesten Mittelwert der Jahresfortgangsnoten aller Klassen der Untersuchung und mit 5,3 den schlechtesten Mittelwert der Lehrerbeurteilungen.
Den zweitschlechtesten Notendurchschnitt, 3,78 bei den Jahresfortgangsnoten und 3,83 bei der Lehrerbeurteilung, hatte die Realschulklasse 9f. Auch sie fiel mit ihren Leistungen teilweise (CT I und II) unter die Leistung der 8. Klasse RS.

Anordnung und relative Unterschiede der CT-Mittelwerte lassen zwei Tendenzen erkennen: sie sind verhältnismäßig konstant von Test zu Test und decken sich mit den Erwartungen, die der Test erfüllen soll.
Während das eine für die Reliabilität spricht, bestätigt das zweite die Validität: der Test mißt offensichtlich, was er messen soll.

5.1.2. Indizes der Testschwierigkeit

Erwartete Leistungsstrukturen wurden auch in Bezug auf die T e x t s c h w i e r i g k e i t von den CT Ergebnissen bestätigt. Es wurde erwartet, daß im allgemeinen niedrigere Rohwerte beim schwierigen Text erreicht werden, als beim leichten, und daß dieses Muster konstant bleibt auf den verschiedenen Leistungsstufen der Lerngruppen.
Hierzu zunächst einige Anmerkungen zur Schwierigkeit der Texte: Es wurde festgestellt, daß CT_1 einen Schwierigkeitsgrad von 1^{st} / 2^{nd} 'grade' hat, CT_2 einen Schwierigkeitsgrad von 2^{nd} 'grade' und CT_3 einen von $2^{nd}/3^{rd}$ 'grade' nach der Fry-Methode. Zur Fry-Methode muß hier eine Anmerkung gemacht werden: ein Faktor ihrer Schwierigkeitsberechnung ist die Satzlänge. Es wird davon ausgegangen, daß bei durchschnittlich längeren Sätzen die Text-

---

24b Es wurde hier bewußt zwischen Test- und Textschwierigkeit unterschieden; während die Testschwierigkeit - errechnet durch ein statistisches Verfahren (Formel 2) - ein Index für die Gesamtzahl richtiger Lösungen ist, ist die Textschwierigkeit eine Angabe zur inhaltlichen ("textlichen") Schwierigkeit.

schwierigkeit wächst. In Text II ist jedoch durch Sätze mit Aufzählungen von Gegenständen die sonst typische kurze Satzlänge des Gesamttextes atypisch verlängert worden, was deshalb in einer nicht unbedingt repräsentativ erfaßten höheren Gesamtschwierigkeit resultiert. Die Textschwierigkeit müßte also etwas nach unten korrigiert werden, was einstimmig von mehreren erfahrenen Englischlehrern bestätigt wurde. Tests I und II dürften somit in ihrer Schwierigkeit als ähnlich eingestuft werden, Test III als der schwierigste der drei Tests.
Die Reaktionen der Klassen auf die unterschiedlichen Textschwierigkeiten wurde nach der Formel 2 berechnet und durch den mittleren Testschwierigkeitsgrad ($\bar{p}$) ausgedrückt.

$$\text{Formel 2: } \bar{p} = \frac{\sum p}{n}$$

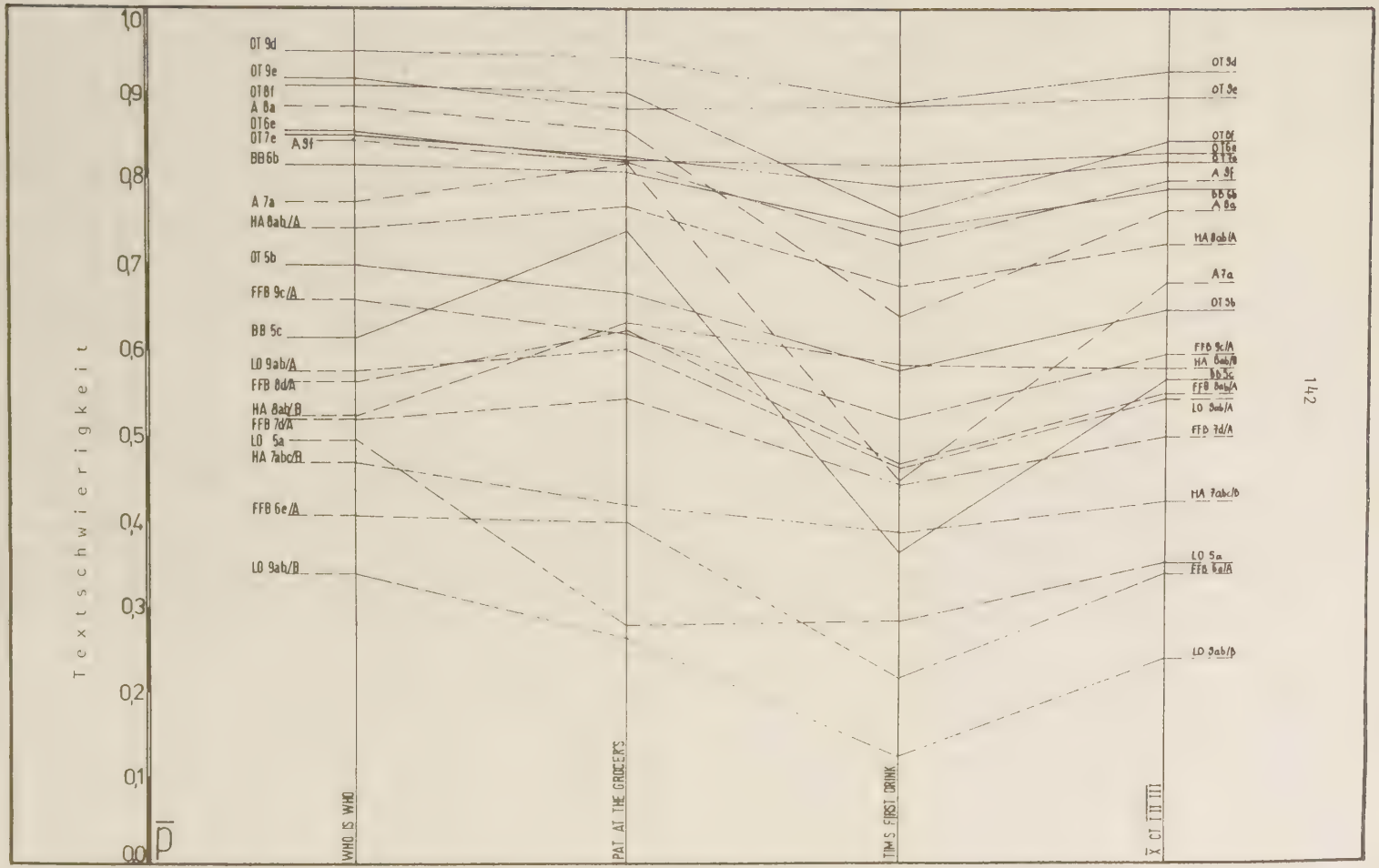
n = Zahl der Items
p = Schwierigkeitsindex = $100.\frac{N_R}{N}$
N_R = Anzahl der Probanden, die die Aufgabe richtig beantwortet haben
N = Gesamtzahl der Probanden

Die Ergebnisse sind in TAFEL 23 zusammengefaßt.
TAFEL 24 zeigt die graphische Repräsentation der Reaktionen der Klassen auf verschiedene Textschwierigkeiten.
Zur Erläuterung sei angmerkt, daß der größte Schwierigkeitsgrad eines Tests, bei dem kein Proband eine Aufgabe löst, durch die Zahl 0, der geringste Schwierigkeitsgrad, wenn alle Probanden alle Aufgaben lösen, durch die Zahl 1 ausgedrückt wird. 0,5 ist ein mittlerer Schwierigkeitsgrad.

Aus den Daten ist folgendes ersichtlich: Die Tests I und II werden im allgemeinen als recht ähnlich oder nur mit geringen Unterschieden in ihrer Schwierigkeit empfunden. Test III wird immer als schwerer empfunden als die anderen zwei Tests. Diese Muster werden gelegentlich durchbrochen: Die Unterschiede zwischen dem Schwierigkeitsindex bei CT I und CT II sind relativ stärker in den Klassen LO 5a, BB 5c und HA 8ab B. Dabei ist CT I für BB 5c und HA 8 ab B erheblich schwerer und für LO 5c erheblich leichter als CT II.
Sowohl BB 5c und LO 5c sind Anfängerklassen. Die Klasse BB 5c hatte allerdings einige Monate Unterricht zwischen der Durchführung von CT I und

Textschwierigkeit

TESTPROFILE



TAFEL 23

Mittlere Schwierigkeitsgrade von drei CTs und der Kombination dieser drei Tests bei den Jahrgangsstufen 5-9 aller drei Schularten

Schulart / Jahrgang \ Test	Schule Klasse	N	I $\bar{p}$	II $\bar{p}$	III $\bar{p}$	I II III $\bar{p}$
HS 5	LO 5a	14	0,5	0,28	0,28	0,36
GY 5	BB 5c	29	0,61	0,74	0,37	0,57
GY 5	OT 5b	26	0,70	0,67	0,58	0,65
HS 6	FFB 6e A Kurs	16	0,41	0,40	0,21	0,34
GY 6	BB 6b	26	0,82	0,81	0,74	0,79
GY 6	OT 6e	32	0,86	0,82	0,82	0,83
HS 7	FFB 7d A Kurs	10	0,52	0,55	0,45	0,51
HS 7	HA 7abc B Kurs	19	0,47	0,43	0,39	0,43
RS 7	A 7a	28	0,78	0,82	0,45	0,68
GY 7	OT 7e	30	0,85	0,82	0,79	0,82
HS 8	FFB 8d A Kurs	12	0,57	0,63	0,47	0,55
HS 8	HA 8ab A Kurs	18	0,75	0,77	0,68	0,73
HS 8	HA 8ab B Kurs	18	0,53	0,64	0,59	0,58
RS 8	A 8a	15	0,89	0,86	0,65	0,77
GY 8	OT 8f	28	0,91	0,90	0,75	0,85
HS 9	LO 9ab A Kurs	21	0,58	0,61	0,47	0,55
HS 9	LO 9ab B Kurs	10	0,34	0,21	0,13	0,24
HS 9	FFB 9c A Kurs	10	0,66	0,62	0,52	0,60
RS 9	A 9f	18	0,85	0,82	0,72	0,79
GY 9	OT 9d	31	0,95	0,94	0,89	0,93
GY 9	OT 9e	29	0,91	0,86	0,89	0,90

CT II , während CT I und CT II in LO 5a in kurzen Abständen hintereinander durchgeführt wurden.
In BB 5c konnte also die Fremdsprachenkompetenz durch Unterricht vergrößert werden, was sich sofort in den CT-Ergebnissen ausdrückte.
Eine ähnliche Beobachtung kann zwischen BB 5a und OT 5b bei CT I gemacht werden.
Für die Klasse OT 5b ist CT I wesentlich leichter als für BB 5c. Ein Blick auf die Zeiten der Testdurchführung zeigt, daß OT 5b ein halbes Jahr mehr Unterricht als BB 5c zur Zeit der Durchführung von CT I hatte.
Anscheinend reagiert der CT sehr fein auf dem Anfangsniveau der fremdsprachlichen Kompetenz.

Ähnliches läßt sich beobachten bei den Unterschieden zwischen CT II und CT III. Hier sind jedoch wegen des größeren Schwierigkeitsgrades unter den Tests die Reaktionsunterschiede allgemein größer.
Eine entgegengesetzte Tendenz kann auf den höheren Leistungsstufen beobachtet werden. Die Unterschiede zwischen leichteren und schwierigen Texten flachen ab. Es tritt ein gewisser 'ceiling'-Effekt ein, d.h. sie können nicht viel besser bearbeitet werden.
Deutlich ist das besonders bei den Klassen OT 9d und OT 9e, bei denen die mittleren Schwierigkeitsgrade bei allen drei Tests um 0,9 liegen.
Im Gegensatz dazu haben die drei Tests für die Klasse HS LO 9ab-B mit Schwierigkeitsgraden von $\bar{p}_{CT\ I} = 0{,}34$, $\bar{p}_{CT\ II} = 0{,}27$, $\bar{p}_{CT\ III} = 0{,}13$ und $\bar{p}_{CT\ I,II,III} = 0{,}24$ die Frustrationsgrenze erreicht.

Nach Harris (1969:145) kann ein Test als leicht eingestuft werden, wenn im Durchschnitt zwischen 70% und 90% der Testantworten richtig sind, das entspricht mittleren Schwierigkeitsgraden zwischen 0,7 und 0,9.
Auf unsere Testgruppen übertragen ist der Test CT I für FFB 9c-A, OTB 5b, HA 8ab-A, A 7a BB 6b, OT 6e und A 8a als leicht einzustufen, für die Klassen OT 8f, OT 9e und OT 9d als zu leicht.
Ebenso nach Harris ist ein Test als schwierig einzustufen, wenn er im Durchschnitt zwischen 50% und 70% richtig beantwortet wird.
Bei weniger als 50%, d.h. unter einem mittleren Schwierigkeitsgrad von 0,5, wäre demnach der Test zu schwierig. Dies wäre hier der Fall bei den Klassen HA 7abc-B, FFB 6e-A und LO 9ab-B.
Test CT II wäre als zu leicht einzustufen für OT 9d, und CT III für keine Klasse.

Test CT II wäre als zu schwer einzustufen für die Klassen LO 9ab-B, LO 5a, FFB 6e-A und HA 7ab-B, und CT III für die Klassen LO 9ab-B, FFB 6e-A, LO 5a, BB 5c, HA 7abc-B, FFB 7d-A, A 7a, LO 9ab-A und FFB 8d-A.

Die Gesamtbatterie wäre für die 9. Gymnasialklassen als zu leicht und für HS LO 5a, HS FFB 6e-A, HS HA 7ab-B und HS LO 9ab-B als zu schwer einzustufen.
Zu einem gewissen Grad äußert sich dies in den Streuwerten der Testbatterie, die bei den Klassen mit den höchsten und den niedrigsten Schwierigkeitsgraden sehr gering sind: OT 9d ($\bar{p}$ = 0,93) hat einen 'range' von 19,4 bis 24,7 Punkten und eine Standardabweichung von 0,99. OT 9e ($\bar{p}$ = 0,90) hat einen 'range' von 20,3 bis 24,7 und eine Standardabweichung von 1,159. LO 9ab-B ($\bar{p}$ = 0,24) hat einen 'range' von 4,3 bis 9,3 und eine Standardabweichung von 1,98.

Z u s a m m e n f a s s e n d kann gesagt werden, daß Unterschiede der Textschwierigkeit durch die CTs registriert wurden. Diese Unterschiede wurden besonders fein auf niedrigen Kompetenzstufen, aber kaum mehr bei den höheren Kompetenzstufen der untersuchten Gruppen erfaßt. Die entworfenen Tests entsprechen insgesamt den Kompetenzniveaus der untersuchten Gruppen, müssen aber bei den höchsten und niedrigsten Gruppen mit Einschränkungen betrachtet werden.

5.1.3. Verteilung der Rohpunkte

Zwei Angaben der Verteilung der Rohpunktwerte wurden gemacht: Der 'range' und die Standardabweichungen. Der 'Range' ist eine grobe Angabe für die Breite der Rohwertverteilung und kann einen ersten Eindruck geben wie homogen eine Klasse ist,oder inwiefern diese Homogenität gestört ist. Der 'Range' ist hier als der Umfang der Weite vom minimalsten Rohwert (x_m bis zum höchsten Rohwert (x_{max}) angegeben.
Die Standardabweichung ist das gebräuchlichste Maß der Verteilung ('centr tendency'). Mit ihrer Hilfe kann das Ergebnis des Einzelnen innerhalb der Gruppe verglichen werden, und es können die Testergebnisse von Gruppen ch rakterisiert und miteinander verglichen werden.Je kleiner die Standardabw chung ist, umso homogener ist eine Gruppe im Hinblick auf das vom Test ge

TAFEL 25: Streuwerte (Range und Standardabweichung) bei drei Cloze Tests und der Kombination der drei CTs für die Schülerjahrgänge 5 - 9 aller Schularten.

Test / Schulart Jahrgang	Schule Klasse	N	I $\bar{X}_{CT}$ RANGE	I SD	II $\bar{X}_{CT}$ RANGE	II SD	III $\bar{X}_{CT}$ RANGE	III SD	I+II+III $\bar{X}_{CT\ 123}$ RANGE	I+II+III SD
HS 5	LO 5a	14	12,5 7-19	2,68	7,07	3,97	7,14	1,875	8,9 4,7-11,3	2,196
GY 5	BB 5c	29	15,48 10-21	2,899	18,62 14-22	2,111	9,21 4-17	3,589	14,26 9,6-17,6	2,004
GY 5	OT 5b	26	17,62 7-24	4,78	16,77 6-24	3,386	14,96 8-23	3,679	16,34 7-22,3	3,311
HS 6	FFB 6e A Kurs	16	10,313 6-18	3,98	10,12 4-17	3,556	5,49 3-9	1,861	8,59 4,6-13	2,771
GY 6	BB 6b	26	20,54 16-24	2,083	20,35 14-25	2,279	18,58 14-23	2,626	19,78 15,3-23	1,880
GY 6	OT 6e	32	21,50 14-25	2,436	20,63 16-25	2,610	20,50 17-24	2,016	20,84 17,3-23,3	1,684
HS 7	FFB 7d A Kurs	10	13,0 9-17	2,514	13,7 10-19	3,02	11,2 7-16	2,70	12,64 10-16,6	2,075
HS 7	HA 7abc B Kurs	19	11,84 4-22	5,199	10,63 2-15	4,058	9,79 3-21	4,354	10,73 4-19,3	3,693
RS 7	A 7a	28	19,46 13-23	3,049	20,61 15-25	2,250	11,32 8-20	2,763	17,08 13,6-21,6	2,144
GY 7	OT 7e	30	21,37 8-25	3,429	20,73 13-25	2,377	19,93 13-24	2,243	20,63 12-24,3	5,068
HS 8	FFB 8d A Kurs	12	14,17 10-19	2,98	15,67 11-19	2,309	11,83 9-18	2,552	13,858 12-18,6	1,786
HS 8	HA 8ab A Kurs	18	18,67 15-22	2,029	19,28 14-23	2,372	17,0 13-21	2,449	18,272 14-20,6	1,756
HS 8	HA 8ab B Kurs	18	13,17 11-15	1,465	15,94 10-19	2,182	14,72 9-24	4,309	14,578 11,6-18,6	1,908
RS 8	A 8a	15	22,2 18-24	2,077	21,47 19-24	1,807	16,13 8-22	3,944	19,187 14-22	2,399
GY 8	OT 8f	28	22,82 18-25	1,806	22,57 18-25	1,834	19,04 12-23	2,835	21,268 17,6-24,3	1,79
HS 9	LO 9ab A Kurs	21	14,48 4-23	5,173	15,14 5-22	4,114	11,76 1-20	4,857	13,738 3,3-20,6	4,213
HS 9	LO 9ab B Kurs	10	8,6 3-16	3,836	6,7 1-12	3,683	3,2 0-6	1,814	6,11 4,3-9,3	1,980
HS 9	FFB 9c A Kurs	10	16,6 8-22	5,967	15,6 12-20	3,718	13,1 8-20	5,343	15,05 9,6-20,6	4,659
RS 9	A 9f	18	21,17 12-25	2,936	20,56 18-24	1,723	18,06 8-22	3,298	19,93 13-22,6	2,198
GY 9	OT 9d	31	23,78 19-25	1,175	23,61 19-25	1,476	22,32 19-25	1,301	23,27 19,4-24,7	0,990
GY 9	OT 9e	29	22,97 18-25	1,592	22,14 17-26	1,885	22,21 18-25	1,916	22,45 20,3-24,7	1,159

N = Anzahl der Probanden
$\bar{X}_{ct}$ = arithmetisches Mittel der Rohwerte
SD = Standardabweichung

messene Persönlichkeitsmerkmal.
Für die Erfassung einer Eigenschaft, die in der Gesamtpopulation breit gestreut ist, erwartet man jedoch eine höhere Standardabweichung der Rohwertergebnisse der Gesamtpopulation.
Die Standardabweichung (SD) wurde nach Formel 3 mit n - 1 Gewichtung berechnet. Mit dieser Gewichtung wird sie üblicherweise als Schätzfunktion für Stichproben verwendet.

$$SD = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n^2}\right) \cdot \frac{n}{n-1}}$$

n Anzahl der Meßwerte
x_i i-ter Meßwert der Häufigkeitsverteilung

Standardabweichungen und Ranges aller Tests sind aufgeführt in TAFEL 25. Alle Angaben sind umtransformiert und auf die tatsächlich erhaltenen von ca 25 möglichen Punkte bezogen.

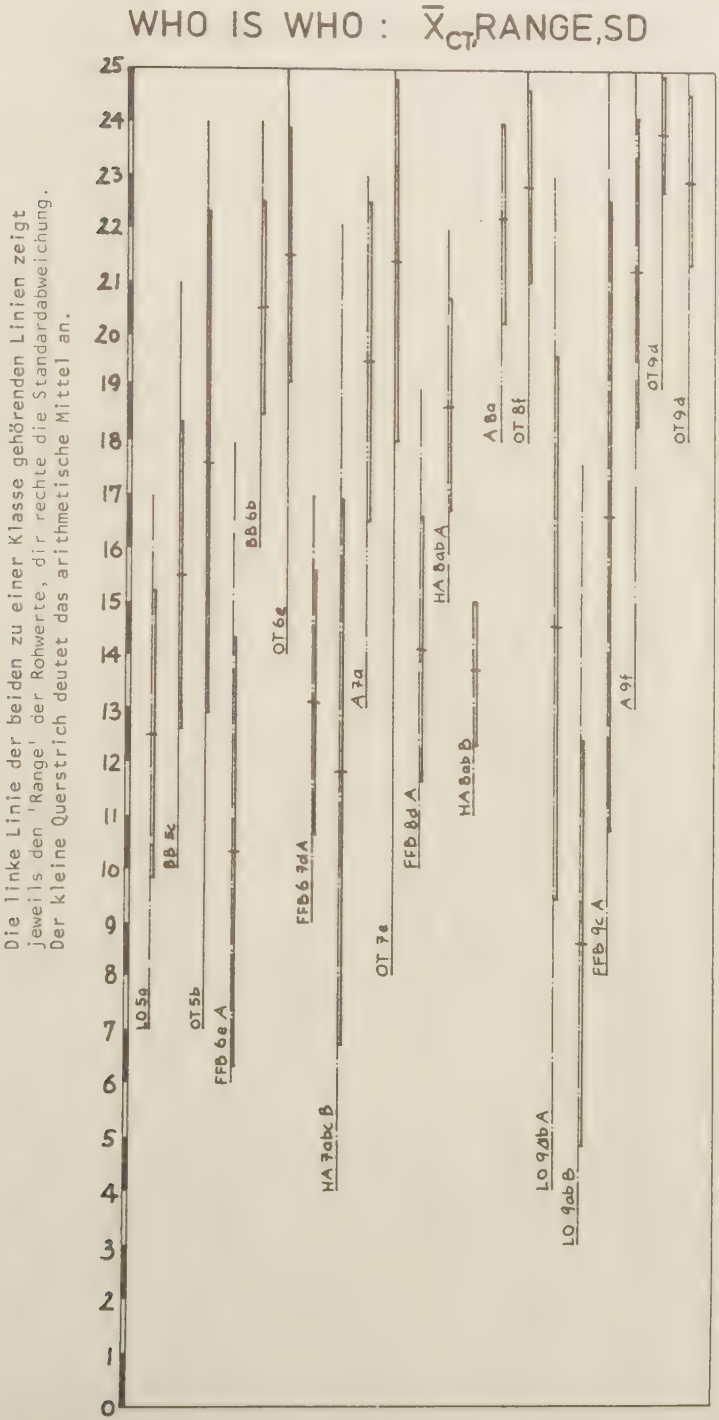
Die TAFELN 26, 27, 28 und 29 zeigen die graphische Repräsentation der Streubreiten ('Ranges') und der Standardabweichungen mit der Angabe des jeweiligen arithmetischen Mittels.
Außer der Standardabweichung wurde auch die Varianz (Var) mit n-Gewichtung berechnet. Die Varianz wurde nach Formel 4 erstellt:

$$Var = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n^2}\right)^2$$

Aus ihr kann gegebenenfalls die Standardabweichung mit n-Gewichtung berechnet werden. Die Varianz ist angegeben in den Listen mit den statistischen Daten der Einzelklassen im Appendix.

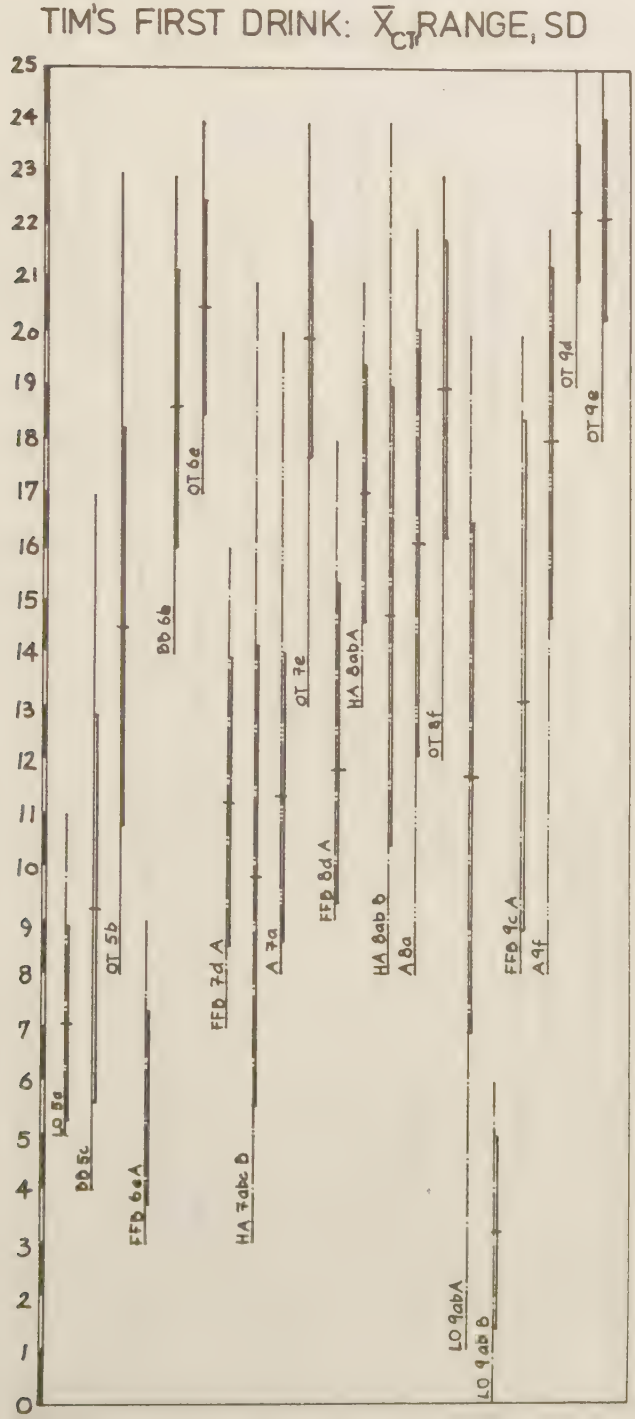
In den graphischen Darstellungen der TAFELN 26, 27, 28 und 29 wird sichtbar, daß die Standardabweichungen besonders bei den niedrigen und mittleren Leistungsstufen größer und bei den höheren Leistungsstufen kleiner sind. Das deutet auf eine größere Homogenität der leistungsstärkeren Gruppen und gleichzeitig auf eine bessere Fähigkeit des Tests, auf den niedrigen und mittleren Leistungsstufen zwischen einzelnen Schülern zu differenzieren, hin.

TAFEL 26: Darstellung der Standardabweichungen und Streubreiten ('Ranges') des CTs "Who is Who" bei den Klassenstufen 5-9 aller Schularten:

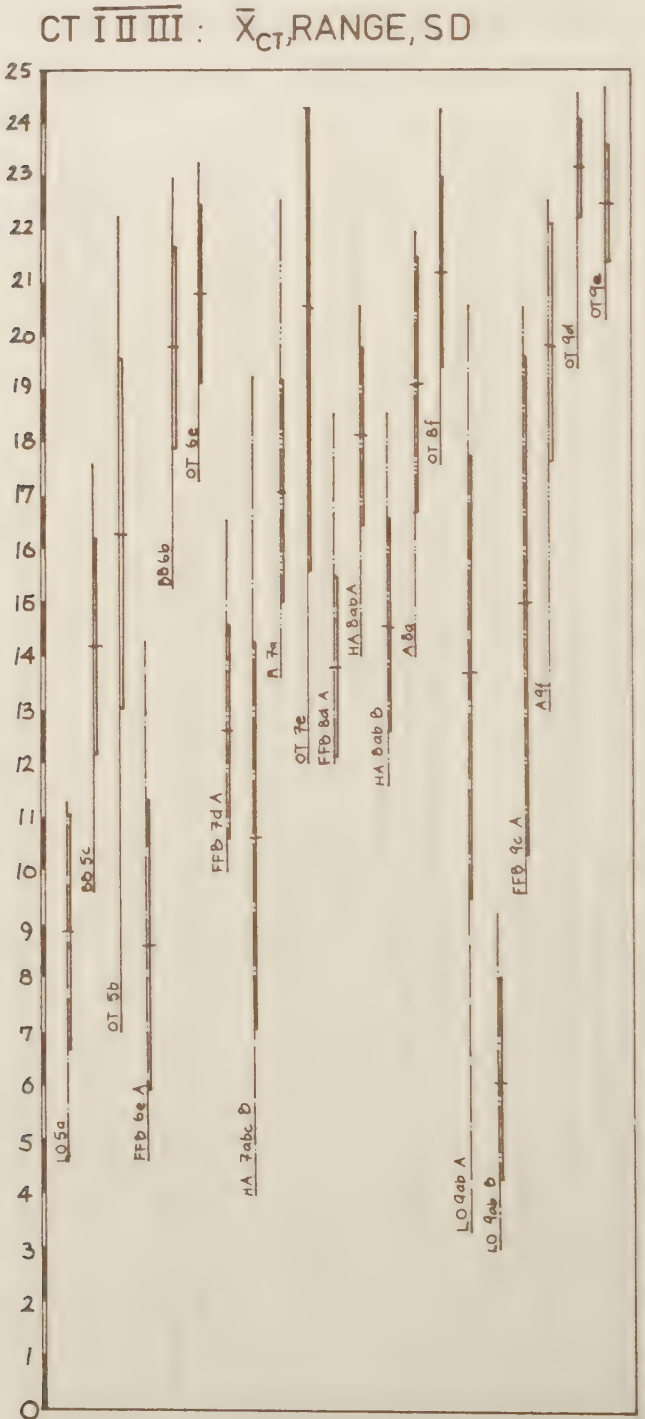


TAFEL 27: Darstellung der Standardabweichungen und Streubreiten des CTs "Pat at the Grocer's" bei den Klassenstufen 5-9 von HS, RS und Gym.:

TAFEL 28: Darstellung der Standardabweichungen und Streubreiten des CTs "Tim's First Drink" bei den Klassenstufen 5-9 von HS, RS und Gym.:



TAFEL 29: Darstellung der Standardabweichungen und Streubreiten der CTs I,II und III bei den Klassenstufen 5-9 aller Schularten:



Die Standardabweichungen bei den einzelnen Tests lagen zwischen 1,175 und 5,367 mit der Mehrzahl der Werte zwischen 2 und 3.
Die Standardabweichungen der Kombination der drei Tests lagen lagen zwischen 0,99 und 5,06. Diese Werte sind relativ niedrig und deuten auf verhältnismäßig homogene Probandengruppen hin.
Die niedrigsten Werte wurden für die Gymnasialklassen 9d und 9e berechnet. (Siehe Tabelle a.a.O.)
Die höchsten Werte wurden für die A-Kurse der neunten Hauptschulklassen und bei der Gesamtbatterie für die Gymnasialklasse OT 7e berechnet.

Die Standardabweichungen der einzelnen Tests (n = 25) bei der Gesamtstichprobe betragen für

CT I = 5,44
CT II = 5,691
CT III = 5,774
CT $\overline{\text{I II III}}$ = 5,032

Sie demonstrieren die Streuung der gemessenen Fähigkeit bei der Gesamtstichprobe und zeigen, daß der CT in akzeptabler Weise in der Lage ist, zwischen den einzelnen Probanden der Gesamtstichprobe zu differenzieren.

Zusammenfassend kann gesagt werden, die Werte der Standardabweichungen weisen die einzelnen Klassengruppen den Erwartungen gemäß als relativ homogene Leistungsgruppen aus. Die Gesamtstreubreite ist groß genug, um einzelne Probanden in akzeptabler Weise voneinander differenzieren zu können.

5.1.4. Die Mittelwerte der Validierungskriterien

Wie bei der Versuchsbeschreibung dargestellt, wurden die Werte zweier Kriterien zur Beurteilung der Fremdsprachenkompetenz der Schüler erhoben. Beide sind in gewissem Sinne Lehrerurteile: einmal wird das Lehrerurteil ausgedrückt in der Jahresfortgangsnote (L), das andere Mal ist es die Einschätzung der "tatsächlichen" Fremdsprachenkompetenz durch den Lehrer ohne die Einwirkung irrelevanter Faktoren, wie dies leider bei der Zensierung geschieht (T). Das letzte sei weiterhin als das Lehrerurteil bezeichnet. Die Mittelwerte jeweils beider Urteile wurden aus den

Rohwerten nach der bereits oben genannten Formel 1 berechnet.
Sie sind zusammengestellt in TAFEL 30. Gleichzeitig ist ihre jeweilige Standardabweichung angegeben.
Zur Erläuterung sei noch einmal angeführt, daß beide Wertungen der allgemein üblichen sechsstufigen Notenskala folgen mit "1" für die beste und "6" für die schlechteste Leistung.bzw. Beurteilung.
Wie bereits ausgeführt, sind die Lehrerbeurteilungen normorientiert, wobei die Norm die Klasse mit dem zu erreichenden Lernziel bildet.
Die Mittelwerte sind deshalb alle konzentriert um den Mittelpunkt der Notenskala und weichen nur geringfügig davon ab. Die größten Abweichungen sind in den A- und B-Kursen der Hauptschulen festzustellen, denn für sie bildet die Gesamtklasse jeweils die Norm, d.h. die besseren Schüler der Gesamtklasse besuchen den A-Kurs, die schlechteren den B-Kurs. Der Durchschnitt aller Jahresfortgangsnoten liegt mit 3,3 etwas über dem Mittelwert von 3,5 der Notenskala.
Im Vergleich mit den Jahresfortgangsnoten sind die Mittelwerte der Lehrerbeurteilungen nahezu konstant zahlenmäßig höher. Der Durchschnitt aller Lehrerbeurteilungen liegt genau auf dem Mittel der Notenskala bei 3,5.
Bei den Lehrerbeurteilungen wird die Fremdsprachenkompetenz der Schüler also insgesamt schlechter beurteilt als bei den Jahresfortgangsnoten, oder die in die Jahresfortgangsnoten einfließenden zur Fremdsprachenkompetenz irrelevanten Faktoren üben einen verbessernden Einfluß aus.
Die Lehrerbeurteilungen weisen insgesamt eine größere Standardabweichung als die Jahresfortgangsnoten auf, d-h. die Streuung der Werte ist breiter. Dies liegt daran, daß Lehrer bei ihren Einschätzungen der Fremdsprachenkompetenz völlig frei und unabhängig waren und auch ohne die Befürchtung schwerwiegender Konsequenzen für die Schüler deren fremdsprachliche Kompetenz mit Extremwerten wie 1, 5 oder 6 bewerten konnten.[24c] Dies ist bei den Jahresnoten nicht immer der Fall, wo oft durch pädagogische Erwägungen die Erteilung von Extremwerten abgeschwächt und verhindert wird.
Insgesamt können die Lehrerbeurteilungen gegenüber den Jahresfortgangsnoten als valider und reliabler betrachtet werden.

24c Selbstverständlich sind eine gewisse Subiektivität und eventuell die Beeinflussung durch den bekannten 'halo-effect' nie ganz auszuschalten.

TAFEL 30: Arithmetische Mittel der Jahresfortgangsnote im Fach Englisch und der Einschätzung zur Beurteilung der fremdsprachlichen Kompetenz von Schülern der Klassen 5-9 aller Schularten:

Schulart / Jahrgang	Schule Klasse	N	$\bar{X}_E$ $\bar{X}_{TE}$	SD_E SD_{TE}
HS 5	LO 5a	14	3,5 3,53	0,65 0,86
GY 5	BB 5c	29	2,93 3,27	0,96 0,8
GY 5	OT 5b	26	2,88 3,29	0,99 1,01
HS 6	FFB 6e AKurs	16	3,5 3,28	0,52 1,14
GY 6	BB 6b	26	3,38 3,40	0,69 0,68
GY 6	OT 6e	32	3,25 3,52	0,916 1,004
HS 7	FFB 7d AKurs	10	3,10 3,40	0,87 1,19
HS 7	HA 7abc BKurs	19	3,58 3,87	0,84 0,83
RS 7	A 7a	28	3,18 3,39	0,61 0,91
GY 7	OT 7e	30	3,27 3,42	1,11 1,23
HS 8	FFB 8d AKurs	12	3,08 3,33	0,51 0,81
HS 8	HA 8ab AKurs	18	2,83 3,25	0,92 0,91
HS 8	HA 8ab BKurs	18	3,94 3,61	0,54 0,96
RS 8	A 8a	15	3,27 3,3	0,70 0,77
GY 8	OT 8f	28	3,21 3,30	0,96 0,92
HS 9	LO 9ab AKurs	21	2,86 3,19	0,72 0,79
HS 9	LO 9ab BKurs	10	4,6 5,3	0,51 0,86
HS 9	FFB 9c AKurs	10	3,1 3,45	0,73 0,38
RS 9	A 9f	18	3,78 3,83	0,65 1,03
GY 9	OT 9d	31	3,22 3,27	0,884 0,93
GY 9	OT 9e	29	3,41 3,31	0,825 1,129

$\bar{X}_E$ Arithmetisches Mittel der Jahresfortgangsnote
$\bar{X}_{TE}$ Arithmetisches Mittel des Lehrerurteils
SD_E Standardabweichung der Jahresfortgangsnote
SD_{TE} Standardabweichung des Lehrerurteils

5.2. Die Reliabilität

Die Reliabilität ist der Grad der Konsistenz mit der ein Test mißt. Sie wird gewöhnlich durch den Reliabilitätskoeffizienten ausgedrückt. Von den verschiedenen Methoden zur Bestimmung der Reliabilität wurde die Split-half Methode mit der Aufwertungsberechnung für die gesamte Testlänge nach der Spearman-Brown-Formel durchgeführt.
Hierzu wurden die Rohwerte aller Tests für die jeweils geradzahligen ('even half') und die jeweils ungeradzähligen ('odd half') Testaufgaben erfaßt. Damit die beiden so entstandenen Testhälften gleich groß waren, wurde die Aufgabe 25 nicht berücksichtigt. Alle Daten sind aufgeführt in den Listen mit den Rohpunktwerten der einzelnen Klassen im Appendix.
Die Korrelationskoeffizienten beider Testhälften (r_{12}) wurden nach der Formel 5 für die Pearson-Produkt-Moment-Korrelation berechnet.

$$r = \frac{m s_x}{s_y} \qquad \text{wobei: } \frac{s_x^2}{s_y^2} = \text{Varianz der } \frac{x}{y}\text{-Werte}$$

Dabei ist s_x^2 = Varianz der x-Werte

s_y^2 = Varianz der y-Werte

$$m = \frac{\frac{\sum x_i \sum y_i}{n} - \sum x_i y_i}{\frac{(\sum x_i)^2}{n} - \sum x_i^2}$$

n = Anzahl der Meßwerte

Aus dem erhaltenen Wert r_{12} wurde mit Hilfe der Spearman-Brown-Formel 6 die Gesamtreliabilität r_{tt} berechnet.

$$r_{tt} = \frac{2 \cdot r_{12}}{1 + r_{12}}$$

Dabei gilt:

r_{12} = Korrelationskoeffizient der zwei Testhälften

Neben dieser Reliabilitätsbestimmung wurde noch wegen seiner größeren Genauigkeit (siehe Kap. 2.1.2.3.) der Reliabilitätskoeffizient nach der Formel von KRISTOF bestimmt ($r_{tt\ KRI}$):

$$r_{tt\ KRI} = \frac{2}{N-1} + \frac{N-3}{N-1} \cdot \frac{4s_1 s_2 r_{12}}{s_1^2 s_2^2 + 2s_1 s_2 r_{12}}$$

Dabei ist

N = Anzahl der Probanden

s_1 / s_2 = Standardabweichung der geraden und ungeraden Hälften

r_{12} = Korrelationskoeffizient der zwei Testhälften

Die Ergebnisse beider Berechnungen sind zusammengestellt in TAFEL 31 für die Einzeltests in den Klassen 5-9 aller Schulstufen.

In den TAFELN 35, 36, 37 und 38 sind die Reliabilitätskoeffizienten neben anderen Werten graphisch repräsentiert. Zur Erläuterung sei darauf verwiesen, daß +1 der höchste Reliabilitätskoeffizient ist und 0 überhaupt keine Konsistenz anzeigt. Die Werte zwischen +1 und 0 zeigen abnehmende Testzuverlässigkeiten an. Sie können allerdings nicht linear, wie z.B. Prozentwerte, miteinander verglichen werden. Linear vergleichbar sind die Werte erst in ihrer quadrierten Form, dem sog. Determinationskoeffizienten, der die gemeinsame Varianz der beiden miteinander korrelierten Größen angibt (siehe auch Kap. 2.1.2.7).

Die Korrelationskoeffizienten nach KRISTOF ($r_{tt\ KRI}$) sind in der Regel etwas besser als die nach SPEARMAN-BROWN berechneten Werte, Durch die nahezu konstantgleichen Abstände zueinander bestätigen sich die beiden Werte gegenseitig.
Da die Methode von Kristof als die präzisere angegeben wird (Lienert, 1967: 223), wird jeweils nur noch auf diese Werte verwiesen.
Die Reliabilitätswerte der Klassen sind bei den einzelnen Tests relativ weit gestreut: bei CT I zwischen 0,443 (FFB 8d-A) und 0,947 mit einem Wert bei -2,284 für HA 8ab-B; bei CT II liegen die Werte mit Schwerpunkten bei 0,7 und 0,8 über die ganze Skala verstreut. Dabei tauchen auch hier bei A 7a mit -0,041 und bei HA 8ab-A mit -0,141 negative Werte auf; bei CT III lie-

TAFEL 31: Reliabilitätskoeffizienten nach SPEARMAN-BROWN und KRISTOF für drei verschiedene CTs und deren Kombination bei den Klassen 5-9 aller Schularten:

Test / Schulart / Jahrgang	Schule Klasse	N	I r_{tt} $r_{tt\,KRi}$	II r_{tt} $r_{tt\,KRi}$	III r_{tt} $r_{tt\,KRi}$	I II III r_{tt} $r_{tt\,KRi}$
HS 5	LO 5a	14	0,519 0,591	0,821 0,808	0,401 0,451	0,771 0,799
GY 5	BB 5c	29	0,559 0,587	0,733 0,224	0,414 0,[illegible]	0,825 0,833
GY 5	OT 5b	26	0,920 0,925	0,695 0,719	0,767 0,784	0,812 0,825
HS 6	FFB 6e A Kurs	16	0,805 0,831	0,836 0,815	-0,782* -0,428	0,717 0,754
GY 6	BB 6b	26	0,479 0,520	0,693 0,711	0,428 0,466	0,610 0,638
GY 6	OT 6e	32	0,724 0,740	0,806 0,814	0,343 0,374	0,606 0,631
HS 7	FFB 7d A Kurs	10	0,715 0,776	0,592 0,682	0,362 0,448	0,598 0,681
HS 7	HA 7abc B Kurs	19	0,910 0,916	0,[illegible] 0,790	0,758 0,782	0,734 0,764
RS 7	A 7a	28	0,688 0,698	-0,130* -0,046	0,458 0,493	0,793 0,797
GY 7	OT 7e	30	0,738 0,755	0,587 0,428	0,511 0,544	0,602 0,630
HS 8	FFB 8d A Kurs	12	0,339 0,443	0,145 0,045	0,543 0,624	0,342 0,441
HS 8	HA 8ab A Kurs	18	0,556 0,605	-0,304* -0,141	0,264 0,346	0,170 0,265
HS 8	HA 8ab B Kurs	18	-1,5[illegible]* -2,284	0,470 0,472	0,670 0,707	0,547 0,596
RS 8	A 8a	15	0,864 0,858	0,628 0,671	0,814 0,841	0,784 0,781
GY 8	OT 8f	28	0,706 0,722	0,689 0,701	0,[illegible] 0,709	0,811 0,816
HS 9	LO 9ab A Kurs	21	0,900 0,907	0,818 0,821	0,850 0,865	0,861 0,874
HS 9	LO 9ab B Kurs	10	0,865 0,865	0,733 0,[illegible]	-0,261* -0,009	0,748 0,803
HS 9	FFB 9c A Kurs	10	0,936 0,947	0,784 0,823	0,876 0,901	0,876 0,903
RS 9	A 9f	18	0,680 0,714	0,199 0,286	0,451 0,515	0,554 0,604
GY 9	OT 9d	31	0,684 0,694	0,586 0,653	-0,367* -0,293	0,438 0,445
GY 9	OT 9e	29	0,414 0,455	0,431 0,440	0,291 0,299	0,476 0,506

gen die Reliabilitätskoeffizienten ebenfalls auf der ganzen Skala verstreut mit negativen Werten für FFB 6e-A bei -0,428, für LO 9ab-A bei -0,009 und für OT 9d bei -0,293.

Die negativen Werte widersprechen dem normalen Charakter des Reliabilitätskoeffizienten, denn sie zeigen an, daß eine "negative" Konsistenz zwischen den beiden Testhälften vorhanden ist. Diese Aussagen müssen als Zufallsergebnis interpretiert werden. Obwohl die erhaltenen Reliabilitätskoeffizienten in der Mehrzahl den Anforderungen von informellen Tests Genüge leisten und zum Teil durch höhere Ansprüche befriedigen, sind sie auf den ersten Blick enttäuschend im Vergleich mit den aus der Literatur bekannten Ergebnissen.

Eine genauere Untersuchung läßt die Ursachen hierfür erkennen:

Aus TAFEL 31 ist ersichtlich, daß keine Konstanz weder innerhalb einzelner Jahrgangsstufen noch Leistungsstufen, Textschwierigkeitsgraden oder Einzeltests zu beobachten ist.

Gleichzeitig lagen die Probandenzahlen dieser Studie - unvermeidbar bedingt durch kleine Klassen und Schülerausfälle - nur zwischen 10 und 31, während die Probandenzahlen der in der Literatur präsentierten Studien zwischen 48 und 442 lagen (siehe TAFEL 16a) und 16b) a.a.O.).

Dazu kommt, daß die Aufgabenanzahl der Studien aus der Literatur mit 50 Items und mehr mindestens doppelt so hoch wie in dieser Studie waren. Man kann schließen, daß die Reliabilität durch den Störfaktor der irrepräsentativen Variabilität und durch den Störfaktor der geringen Aufgabenanzahl negativ beeinflußt war.

Beide Faktoren konnten leicht durch Zusammenfassung der Probandenzahl bzw. der Einzeltests geändert werden. Eine Erhöhung der Probandenzahl und Aufgabenanzahl ergab dann auch in der vorliegenden Studie Ergebnisse, die den anderen Untersuchungen aus der Literatur entsprechen.

Folgende Reliabilitätskoeffizienten wurden für die G e s a m t z a h l d e r P r o b a n d e n N = 440 berechnet:

	CT I	CT II	CT III
r_{tt} SPEARM.-BROWN	0,897	0,878	0,823
r_{tt} KRISTOF	0,897	0,876	0,898

Diese Ergebnisse zeigen unter den Bedingungen der vorliegenden Studie, daß der untersuchte CT ein reliables Meßinstrument ist, und daß die in der Literatur vorgefundenen Ergebnisse auch für die ausgewählte Probandengruppe mit niedrigem Kompetenzniveau bestätigt werden.

Durch die Zusammenfassung der drei Einzeltests wurde eine E r h ö h u n g d e r A u f g a b e n z a h l auf 75 erreicht. Die Ergebnisse sind zusammengestellt in der Spalte CT I II III der TAFEL 31. Für die einzelnen Klassen sind durch die Erhöhung der Aufgabenzahl die Reliabilitätskoeffizienten trotz kleiner Probandenzahlen insgesamt höher als bei den Einzeltests. Bis auf vier Werte liegen alle Werte im Bereich um 0,6 oder darüber. Keine Werte liegen im negativen Bereich, dabei sind die Werte von OT 9d und OT 9e mit 0,445 bzw. 0,506 zweifellos durch die schlechte Validierung der Testschwierigkeit für diese Klassen beeinflußt.
Für die vorliegende Probandengruppe kann die Feststellung von Rand (1978; siehe Kapitel 2.4.3.), daß eine Testlänge von 25 Aufgaben genügt, nicht bestätigt werden. Bei der vorliegenden Untersuchung konnte durch eine empirische Testverlängerung von 25 auf 75 Aufgaben eine deutliche Verbesserung der Reliabilitätskoeffizienten beobachtet werden.
Neben der empirischen Testverlängerung gibt es auch die Möglichkeit, die Steigerung der Reliabilität durch eine Verlängerung des Tests mathematisch zu berechnen. Hierzu dient eine Formel von SPEARMAN-BROWN (in Lienert,[2] 1967: 243). Formel 8:

$$r'_{tt} = \frac{\frac{n'}{n} \cdot r_{tt}}{1 + (\frac{n'}{n} - 1) \cdot r_{tt}}$$

In dieser Formel bedeuten:

r'_{tt}	voraussichtliche Reliabilität des veränderten Tests
r_{tt}	Reliabilität des ursprünglichen Tests
n	Anzahl der Aufgaben im ursprünglichen Test
n'	Anzahl der Aufgaben im veränderten Test

Um die allgemein zur Vergleichbarkeit von Reliabilitätskoeffizienten übliche Testlänge von 100 Aufgaben zu erhalten, wurden die Reliabilitätskoeffizienten für das Vierfache der ursprünglichen Aufgabenzahl berechnet ($r_{tt\ 100}$). Sie sind zusammengefaßt in TAFEL 32, wobei $r_{tt\ 100}$ bedeutet: Reliabilitätskoeffizient berechnet nach SPEARMAN-BROWN für 100 Aufgaben für drei CTs mit 25 Aufgaben und deren Kombination in den Jahrgangsstufen 5-9 aller Schularten.

Die Ergebnisse für $r_{tt\ 100}$ für die Gesamtpopulation N = 440 sind:

	CT I	CT II	CT III
$r_{tt\ 100}$	0,972	0,967	0,972

Insgesamt kann der untersuchte CT als reliables Meßinstrument betrachtet werden und somit die Ansprüche der Teststandardisierung befriedigen.

5.3. V a l i d i t ä t

Die statistische Definition der empirischen oder kriteriumsbezogenen Validität ist ein nachweisbarer Zusammenhang zwischen den Rohpunktwerten des Tests und einem als valide erkannten Kriterium. Zur Bestimmung der äußeren Validität wurden die Testergebnisse mit jeweils zwei Werten der Lehrerbeurteilung korreliert. Diese zwei Werte sind die Jahresfortgangsnote in Englisch (E) und die Lehrerbeurteilung (TE). Von diesen beiden Werten wurde der letztere als die genauere und gültigere erkannt.
Die Korrelationskoeffizienten wurden berechnet als PEARSON-PRODUKT-MOMENT-KORRELATION nach der Formel 5, die auch zur Berechnung der Reliabilitätskoeffizienten verwendet wurde. Die Ergebnisse sind zusammengestellt in TAFEL 33.
Um die Bedeutung der Beziehungen zwischen den Testergebnissen und den Kriterien zu demonstrieren, wurden die jeweiligen Stufen der Signifikanz angegeben. (Nach Fisher & Yates, "Statistical tables for biological, agricultural and medical research," Brown & Amos & Mink, ^{2}1975:135).
Zur Erläuterung sei angemerkt, daß vorhandene Beziehungen zwischen den CT-

TAFEL 32: Reliabilitätskoeffizient berechnet nach SPEARMAN-BROWN für 100 Aufgaben für drei Cloze Tests mit 25 Aufgaben und deren Kombination in den Jahrgangsstufen 5-9 aller Schularten:
(*Für diese Werte ist Formel 8 nicht anwendbar)

Test / Schulart / Jahrgang	Schule Klasse	N	I $r_{tt\ 100}$	II $r_{tt\ 100}$	III $r_{tt\ 100}$	I II III $r_{tt\ 100}$
HS 5	LO 5a	14	0,853	0,944	0,756	0,941
GY 5	BB 5c	29	0,851	0,536	0,803	0,952
GY 5	OT 5b	26	0,747	0,91	0,736	0,950
HS 6	FFB 6e A Kurs	16	0,952	0,946	6,028*	0,925
GY 6	BB 6b	26	0,813	0,908	0,772	0,876
GY 6	OT 6e	30	0,919	0,946	0,705	0,872
HS 7	FFB 7a A Kurs	10	0,933	0,826	0,765	0,895
HS 7	HA 7abc B Kurs	19	0,978	0,938	0,935	0,928
RS 7	A 7a	28	0,902	-0,213*	0,799	0,940
GY 7	OT 7e	30	0,923	0,950	0,827	0,971
HS 8	FFB 8d A Kurs	12	0,761	0,159	0,869	0,759
HS 8	HA 8ab A Kurs	18	0,860	-0,977*	0,679	0,591
HS 8	HA 8ab B Kurs	18	1,531*	0,781	0,906	0,855
RS 8	A 8a	15	0,960	0,841	0,955	0,934
GY 8	OT 8f	28	0,912	0,902	0,907	0,947
HS 9	LO 9ab A Kurs	21	0,975	0,948	0,962	0,965
HS 9	LO 9ab B Kurs	10	0,962	0,939	-0,036*	0,943
HS 9	FFB 9c A Kurs	10	0,986	0,949	0,973	0,974
RS 9	A 9f	18	0,909	0,616	0,809	0,859
GY 9	OT 9d	31	0,900	0,883	-9,681*	0,762
GY 9	OT 9e	29	0,770	0,759	0,630	0,804

FEL 33: Korrelationskoeffizienten zwischen den Rohpunktwerten von drei CTs und deren Kombination und zwei Validierungskriterien für die Jahrgangsstufen 5-9 bei allen Schularten:

(Die Hochzahlen geben das Signifikanzniveau in Prozenten an.)

Test / Schulart / Jahrgang	Schule Klasse	N	I r CT/E r CT/TE	II r CT/E r CT/TE	III r CT/E r CT/TE	I II III $r_{CT123/E}$	 $r_{CT123/TE}$
HS 5	LO 5a	14	-0,595[5] -0,539[5]	-0,074 -0,695[1]	-0,063 -0,572[5]	-0,304	-0,798[1]
GY 5	BB 5c	29	-0,526[1] -0,576[1]	-0,133 -0,376[10]	-0,089 -0,325[10]	-0,361[10]	-0,623[1]
GY 5	OT 5b	26	-0,607[1] -0,906[1]	-0,555[1] -0,750[1]	-0,280 -0,628[1]	-0,599[1]	-0,909[1]
HS 6	FFB 6e A Kurs	16	-0,600[2] -0,785[1]	-0,617[2] -0,807[1]	-0,104 -0,565[5]	-0,575[2]	-0,847[1]
GY 6	BB 6b	26	-0,093 -0,450[4]	-0,163 -0,560[1]	-0,519[1] -0,798[1]	-0,343[10]	-0,788[1]
GY 6	OT 6e	32	-0,564[1] -0,616[1]	-0,592[1] -0,761[1]	-0,454[1] -0,634[1]	-0,764[1]	-0,946[1]
HS 7	FFB 7d A Kurs	10	-0,510 -0,823[1]	-0,329 -0,455	-0,479 -0,612[10]	-0,565[10]	-0,812[1]
HS 7	HA 7abc B Kurs	19	-0,488[5] -0,563[2]	-0,457[5] -0,806[1]	-0,361 -0,231	-0,543[2]	-0,658[1]
RS 7	A 7a	28	-0,384[3] -0,644[1]	-0,270 -0,461[2]	-0,583[1] -0,666[1]	-0,537[1]	-0,760[1]
GY 7	OT 7e	30	-0,506[1] -0,809[1]	-0,533[1] -0,785[1]	-0,352[6] -0,638[1]	-0,549[1]	-0,887[1]
HS 8	FFB 8d A Kurs	12	-0,724[1] -0,325	-0,025 -0,496[10]	-0,542[10] -0,478	-0,579[5]	-0,595[5]
HS 8	HA 8ab A Kurs	18	-0,659[1] -0,668[1]	-0,139 -0,619[1]	-0,052 -0,566[2]	-0,319	-0,767[1]
HS 8	HA 8ab B Kurs	18	+0,161 -0,139	+0,047 -0,696[1]	-0,235[1] -0,364[1]	-0,127	-0,946[1]
RS 8	A 8a	15	-0,430 -0,484[10]	-0,610[2] -0,541[5]	-0,606[2] -0,762[1]	-0,617[2]	-0,743[1]
GY 8	OT 8f	28	-0,148 -0,145	-0,389[5] -0,327[10]	-0,481[2] -0,482[2]	-0,519[1]	-0,490[1]
HS 9	LO 9ab A Kurs	21	-0,513[2] -0,677[1]	-0,528[2] -0,747[1]	-0,279 -0,742[1]	-0,486[5]	-0,804[1]
HS 9	LO 9ab B Kurs	10	-0,314 -0,839[1]	-0,117 -0,638[5]	-0,380 -0,150	-0,343	-0,978[1]
HS 9	FFB 9c A Kurs	10	-0,570[10] -0,777[1]	-0,470 -0,944[1]	-0,623[10] -0,916[1]	-0,606[10]	-0,933[1]
RS 9	A 9f	18	-0,692[1] -0,678[1]	-0,252 -0,376	-0,490[5] -0,673	-0,62[1]	-0,73[1]
GY 9	OT 9d	31	-0,110 -0,033	-0,314[10] -0,224	-0,094 -0,131	-0,157	-0,117
GY 9	OT 9e	29	+0,066 -0,369[5]	+0,169 -0,298[10]	-0,101 -0,336[5]	0,064	-0,521[1]

Rohwerten und den Werten der Lehrerbeurteilung in der Regel durch negative Korrelationskoeffizienten ausgedrückt werden. Das liegt daran, daß die Beurteilung der höchsten Kompetenz durch den niedersten Skalenwert = 1 und die niedrigste Kompetenz durch den höchsten Skalenwert = 6 ausgedrückt werden. Im Gegensatz dazu liefert der CT die höchsten Punktwerte = 25 bei vollständiger Testlösung und den niedrigsten Punktwert = 0 bei völligem Versagen. Eine perfekte Korrelation zwischen den beiden Werten wird deshalb durch den Korrelationskoeffizienten r = -1 ausgedrückt. Bei 0 besteht keine Korrelation. Die zwischen 0 und -1 liegenden Werte können nicht linear miteinander verglichen werden, sondern nur in ihrer Quadratur als Determinationskoeffizienten.
Die Angabe des Signifikanzniveaus sagt aus, wie groß die Wahrscheinlichkeit ist, daß eine Korrelation auf Zufall beruht. Auf dem 0,01 oder 1%-Niveau besteht die aufgezeigte Beziehung zu einem Prozent auf Zufall, auf dem 0,5 oder 5%-Niveau besteht die Möglichkeit einer zufälligen Übereinstimmung zu 5%. Eine Aussage auf den 0,01 oder 1%-Niveau ließe also auf eine größere Zuverlässigkeit schließen als auf dem 5%-Niveau.

Zunächst ein Blick auf beide Korrelationskoeffizienten. Die Korrelation der CT-Rohwerte mit dem jeweiligen Lehrerurteil ($r_{ct\ /\ TE}$) ist mit geringfügigen Ausnahmen bei Einzeltests höher als die Korrelation der CT-Ergebnisse mit den Jahresfortgangsnoten ($r_{CT\ /\ E}$). In HS 8 FFB 8d-A ist die Korrelation sowohl für CT I als auch für CT III für $r_{CT/TE}$ geringer als für $r_{CT/E}$. Eine Nachforschung ergab, daß die Lehrerin die Klasse erst vor kurzem übernommen hatte.
Auch die Signifikanz von $r_{CT/TE}$ ist entweder gleich oder besser als von $r_{CT/E}$. Es kann daraus geschlossen werden, daß der CT eher ein Meßinstrument der von den Lehrern geschätzten "tatsächlichen" Fremdsprachenkompetenz ist als ein Meßinstrument der durch die Jahresnote ausgedrückten "Schulkompetenz".
Die Werte von $r_{CT/TE}$ liegen im allgemeinen bei den Einzeltests um mindestens 0,5 oder darüber; nur in Einzelfällen liegen sie darunter. Betrachtet man $r_{CT/TE}$ bei der Kombination CT I II III so sind die Werte wesentlich höher und auf dem 1%-Niveau signifikant bis auf $r_{CT/E}$ von OT 9d, der mit -0,117 nicht signifikant, und $r_{CT/E}$ von HS FFB 8d-A, der mit -0,595 auf dem 5%-Niveau signifikant ist. Aus den nahezu konstant höheren Validitätskoeffizienten bei der Kombination der drei Tests gegenüber den Einzeltests kann auf die

Irrepräsentabilität der Einzeltests geschlossen werden. Dies wird auch bestätigt durch die Tatsache, daß keine konstanten Interkorrelationen der Einzeltests auf den verschiedenen Klassenstufen nachweisbar sind. Siehe TAFEL 34.

Die Validitätskoeffizienten ($r_{CT123/TE}$) liegen, außer bei OT 9d immer um -0,5 oder in der Mehrzahl der Klassen wesentlich darüber.[25]
Nach Lienert können diese Werte durchaus als akzeptabel gesehen werden (siehe Kap. 2.1.2.4.), um dem Test eine ausreichende Validität im Hinblick auf das vom Kriterium gemessene Persönlichkeitsmerkmal der untersuchten Probandengruppe zuzuerkennen.

Auch in der vorgestellten Literatur wurden ähnliche Korrelationswerte (zwischen 0,41 und 0,90) als akzeptabel zur Validierung des CT bei den jeweiligen Probandengruppen gesehen.

5.4. Ökonomie und Testadministration

Bei der Testdurchführung wurden die in Kapitel 2.1.2.6. aufgestellten Bedingungen beobachtet, die ein ökonomischer Test erfüllen soll. Die reine Testzeit betrug je nach Leistungsstufe zwischen ca 10[25a] und 25 Minuten. Mit Vor- und Nachbesprechung konnte ein Test jeweils leicht innerhalb einer Schulstunde durchgeführt werden. Bei der Durchführung eines längeren Tests von ca 75 bis 100 Items verlängert sich diese Zeit nicht um das dreifache, sondern verdoppelt sich höchstens, da die organisatorischen Teile der Testdurchführung nur einmal zu erledigen sind. Im Vergleich mit anderen Standardisierten Tests, bei denen Durchführungszeiten von bis zu mehreren Stunden zu verzeichnen sind, (z.B. HET 6-), sind diese Ergebnisse sehr vorteilhaft. Ebenso vorteilhaft ist der geringe Materialaufwand von einer DIN A4-Seite pro ca 25 Aufgaben gegenüber den traditionellen Testheften mit separaten Antwortblättern.
Zur Durchführung bedarf es nur einer einfachen Erklärung,des Testblattes und keiner weiteren Medien. Die Testbögen können ohne Schwierigkeit ausgefüllt werden. Es hat sich allerdings gezeigt, daß verschiedene Schüler es als vorteilhaft empfanden, die Lücken im Text zuerst auszufüllen und erst bei der abschließenden Überprüfung die Antworten in die Zeilen am Rand einzutragen. Dadurch konnte der Text vom Schüler beim Durchlesen noch

---

25a Dies trifft insbesondere für die 9. Klassen des Gymnasiums zu, für die die Tests als sehr leicht erkannt wurden.

TAFEL 34: Interkorrelation der drei Cloze Tests bei den Jahrgangsstufen 5-9 von HS, RS und GYM.:

$r_{CT1/CT2}$ = Korrelation zwischen CT_1 und CT_2

Schulart Jahrgang	Schule Klasse	N	$r_{CT1/CT2}$	$r_{CT1/CT3}$	$r_{CT2/CT3}$
HS 5	LO 5a	14	0,365	0,168	0,515
GY 5	BB 5c	29	0,244	0,107	0,416
GY 5	OT 5b	26	0,693	0,558	0,417
HS 6	FFB 6e A Kurs	6	0,849	0,584	0,656
GY 6	BB 6b	26	0,532	0,527	0,540
GY 6	OT 6e	32	0,307	0,053	0,408
HS 7	FFB 7d A Kurs	10	0,048	0,455	0,58
HS 7	HA 7abc B Kurs	4	0,585	0,508	0,37
RS 7	7a	28	0,443	0,504	0,416
GY 7	OT 7e	30	0,647	0,418	0,45c
HS 8	FFB 8a A Kurs	12	0,122	0,493	0,298
HS 8	HA 8ab A Kurs	18	0,607	0,260	0,547
HS 8	HA 8ab B Kurs	18	-0,163	-0,085	0,517
RS 8	A 8a	31	0,415	0,221	0,413
GY 8	OT 8f	28	0,138	0,124	0,591
HS 9	LO 9ab A Kurs	21	0,744	0,608	0,775
HS 9	LO 9ab B Kurs	10	0,289	0,013	-0,323
HS 9	FFB 9c A Kurs	10	0,793	0,814	0,825
RS 9	A 9f	18	0,399	0,698	0,305
GY 9	OT 9d	31	0,352	0,006	0,293
GY 9	OT 9e	29	0,240	0,002	0,071

leichter im Gesamtüberblick erfaßt werden und Verfahrensfehler, wie etwa falsches Eintragen in die Antwortspalten, reduziert werden.
Bei der Gruppentestadministration wurde wie oben erwähnt gerade die Übersichtlichkeit der Antworten als Nachteil empfunden, da benachbarte Schüler leicht zum gegenseitigen Abschreiben verleitet wurden. Diese Tatsache dürfte gelegentlich Testergebnisse verfälscht und die Validitätskoeffizienten negativ beeinflußt haben. Abhilfe durch strengere Aufsicht und getrennte Sitzweise ist dringend anzuraten.

5.5. Interdependenz der statistischen Werte

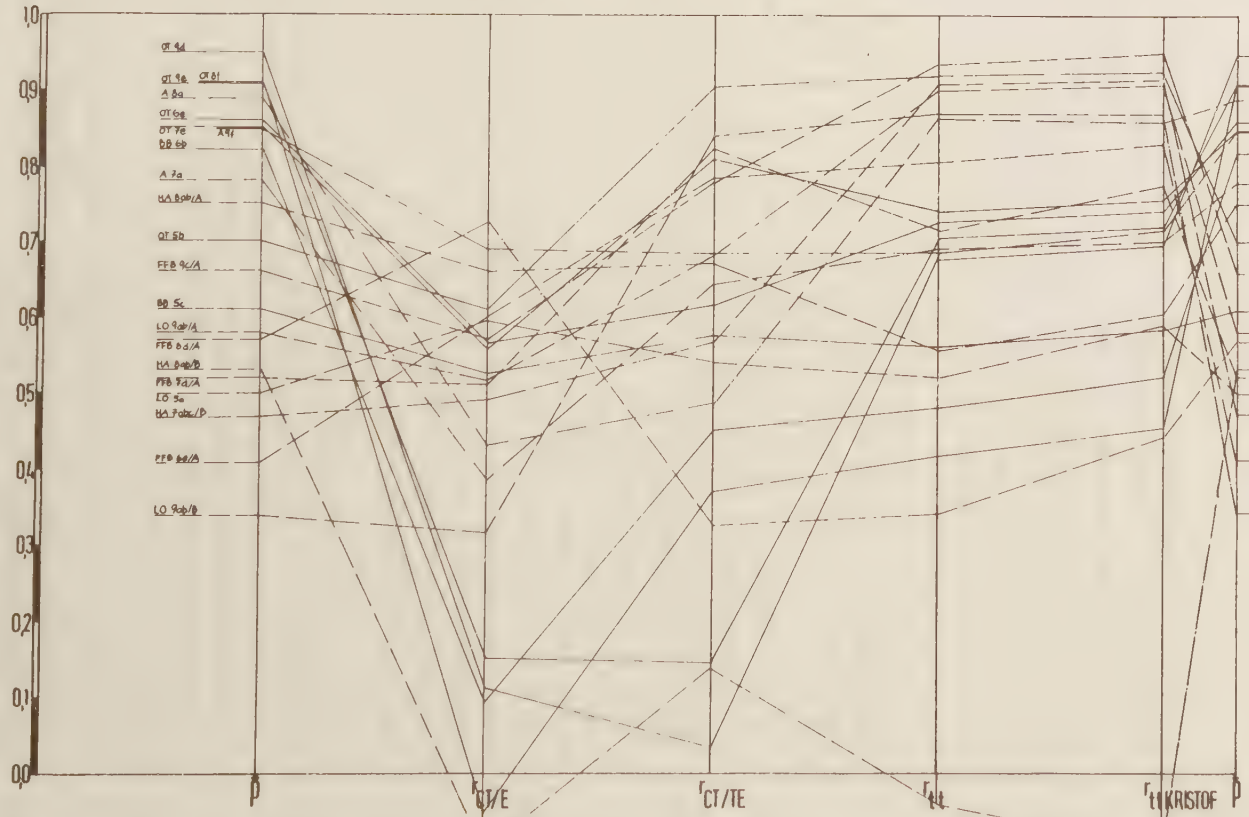
Die TAFELN 35, 36, 37, und 38 zeigen bei den drei CTs und bei der Kombination der drei Tests die Verteilung des durchschnittlichen Schwierigkeitsgrades der Tests (p), der Korrelation zwischen dem CT und der Englischnote respektive dem Lehrerurteil und die Verteilung der beiden Reliabilitätskoeffizienten im Vergleich miteinander, bzw. graphisch nebeneinanderstehend für jede einzelne Klasse.
Auffällig ist zunächst, daß die Muster der Verteilung in allen drei Tests und in der Kombination aller drei Tests deutlich ähnlich sind. Die Rangordnung der verschiedenen Klassen ist bei den einzelnen Test bei allen Werten nicht wesentlich unterschiedlich. Die Aussage, daß die verschiedenen Tests das gleiche Persönlichkeitsmerkmal mit der gleichen Güte messen, kann dadurch bekräftigt werden.

Bereits hingewiesen wurde auf das Verhältnis zwischen r_{tt} und $r_{tt\ KRI}$. Die Graphik zeigt, daß $r_{tt\ KRI}$ nahezu konstant besser ist als r_{tt}. Ebenso wurde bereits der konstante Unterschied zwischen $r_{CT/E}$ und $r_{CT/TE}$ erwähnt. Die beträchtliche höheren Werte für $r_{CT/TE}$ in der großen Mehrzahl der Fälle sind klar ersichtlich und zeugen für seine größere Validität in Bezug auf das Kriterium des Lehrerurteils.

Beachtenwert sind die Zusammenhänge zwischen den Testschwierigkeitsgraden und den Validitätskoeffizienten, bzw. den Reliabilitätskoeffizienten.
Die Muster sind ähnlich für alle drei Tests, bzw. deren Kombination.

Die Klassen mit den numerisch höchsten Schwierigkeitsgraden weisen die

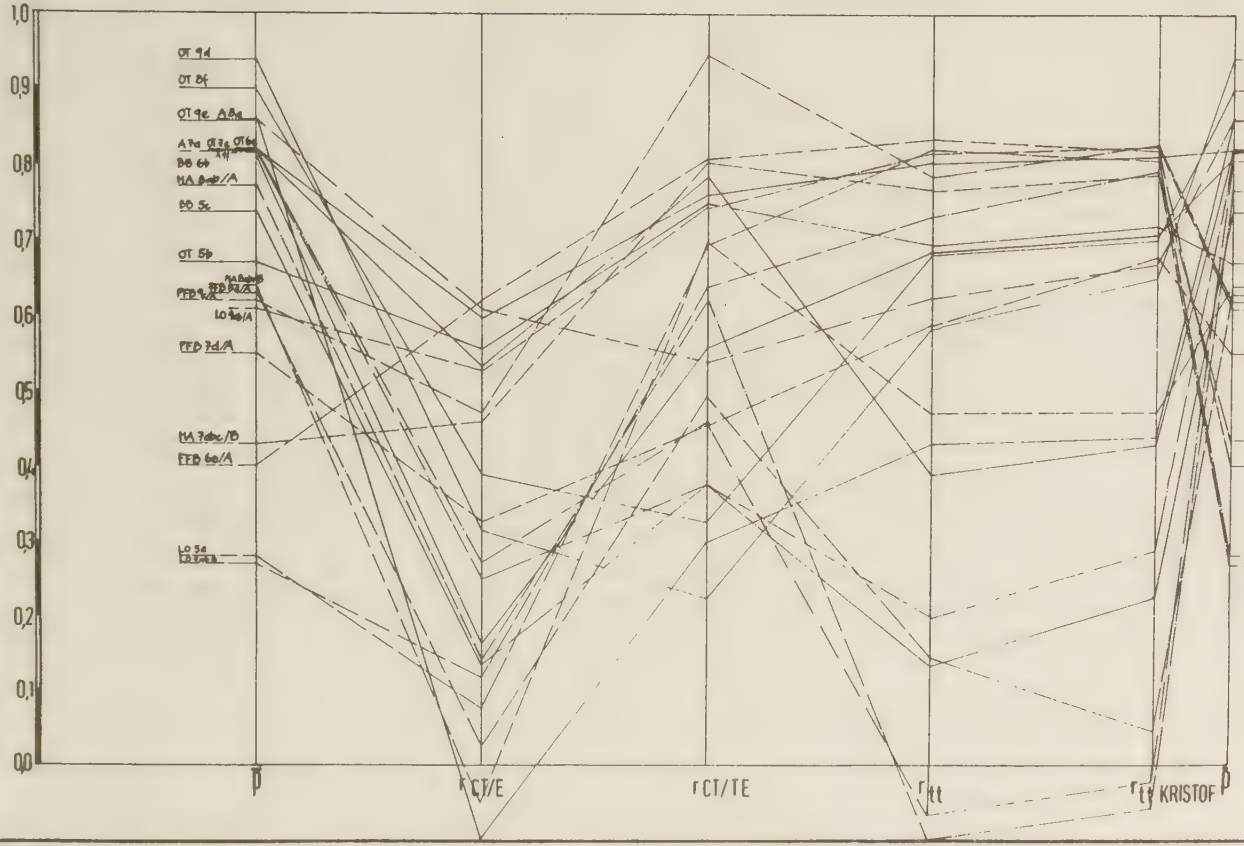
KORRELATIONSPROFILE: WHO IS WHO



TAFEL 35

Zur Interdependenz der statistischen Werte bei CT I und der Kombination dieser Ts

KORRELATIONSPROFILE: PAT AT THE GROCERS

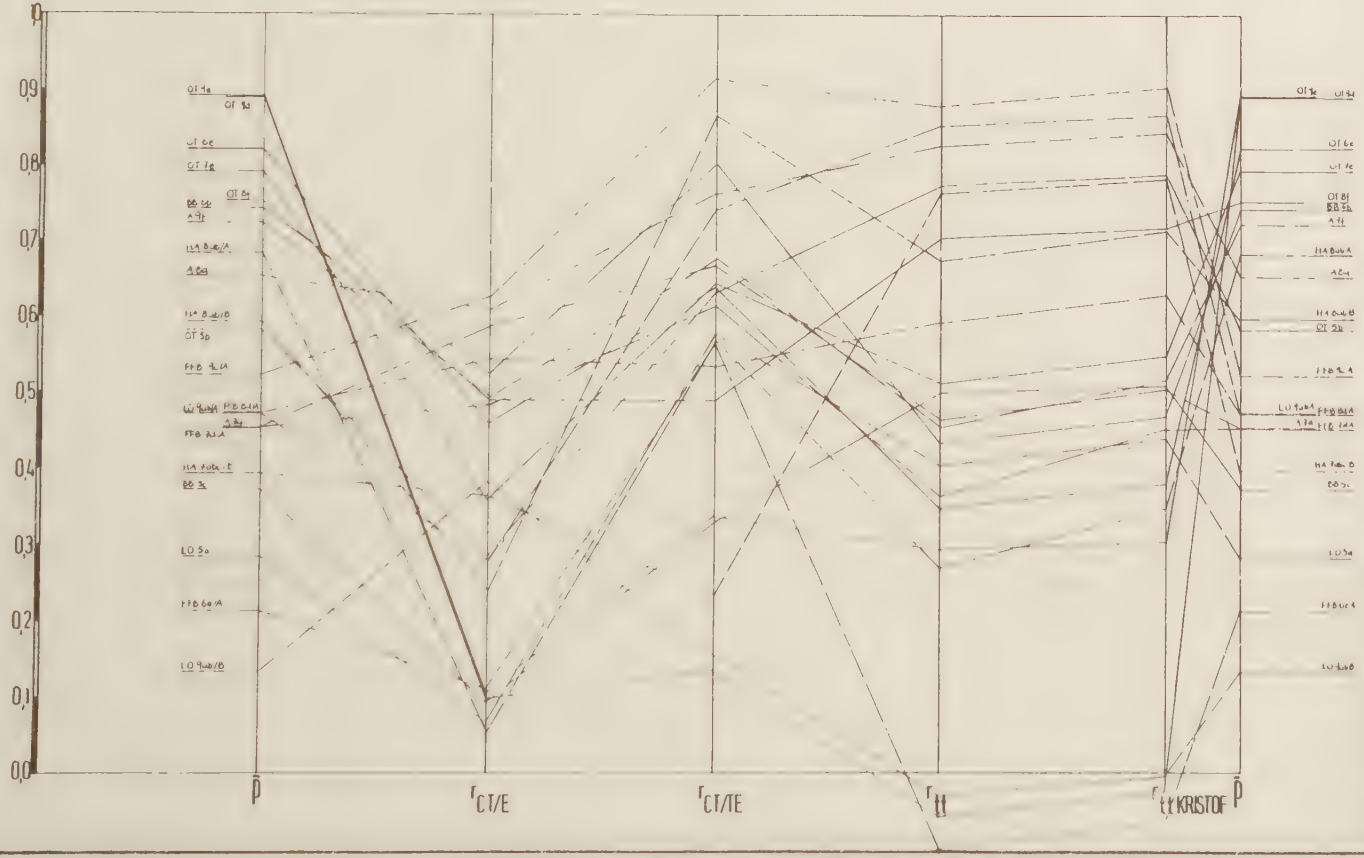


Zur Interdependenz der statistischen Werte bei CT Nummer 11 und der Kombination dieser Tests:

TAFEL 36

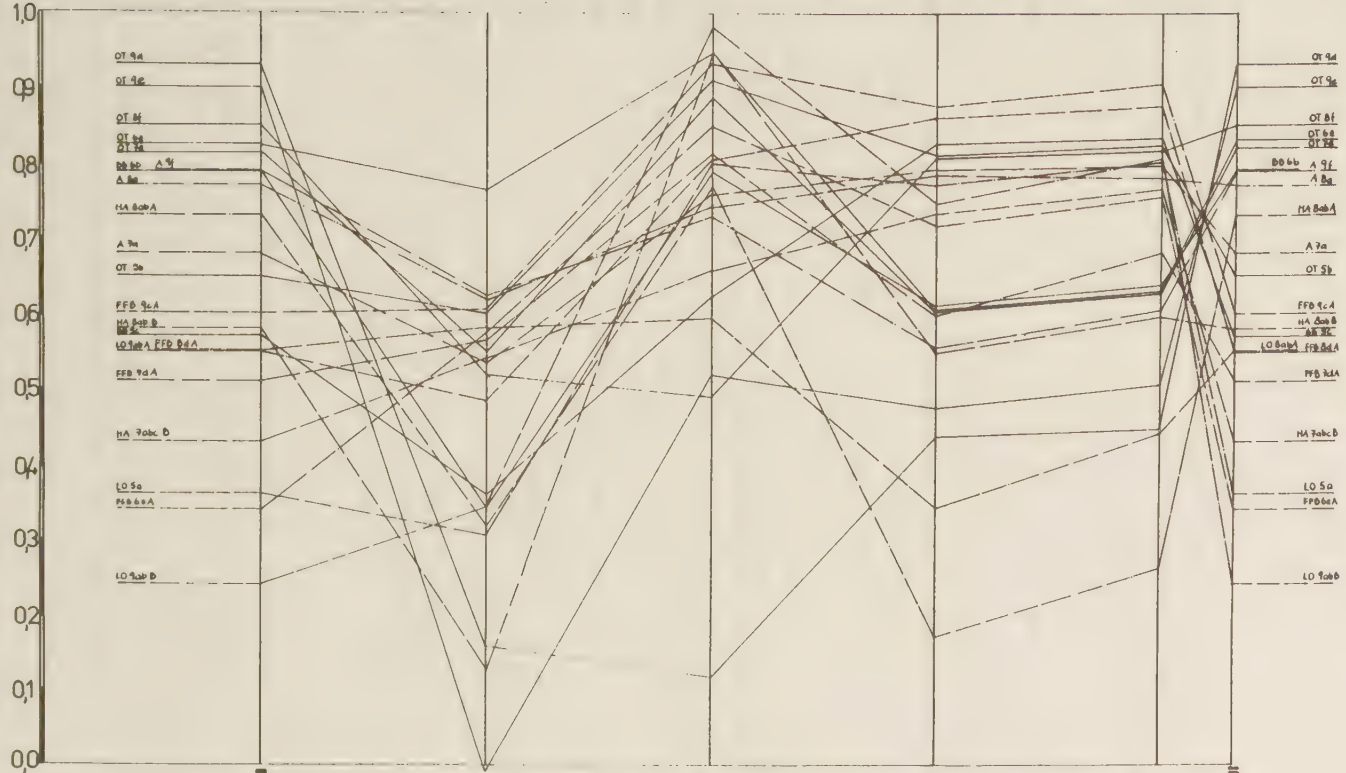
KORRELATIONSPROFILE: TIM'S FIRST DRINK

Zur Interdependenz der statistischen Werte bei CT III und der Kombination dieser drei Tests der Jahrgangsstufen 5-9:



KORRELATIONSPROFILE: CT I II III

Zur Interdependenz der statistischen Werte bei der Kombination der drei Tests CT I, II und III der Jahrgangsstufen 5-9 aller Schularten:



TAFEL 38

niedrigsten Validitätskoeffizienten mit dem Kriterium der Jahresfortgangsnote auf. Bei numerisch niedrigen Schwierigkeitsgraden steigen diese Validitätskoeffizienten in der Regel.
Die Korrelation mit dem Kriterium des Lehrerurteils liegt parallel dazu jeweils etwas höher.
Ähnliche Beobachtungen können für die Beziehungen zwischen den Reliabilitätskoeffizienten und den Schwierigkeitsgraden gemacht werden. Klassen mit numerisch hohen Schwierigkeitsgraden haben niedrigere, und Klassen mit numerisch niedrigen Schwierigkeitsgraden haben höhere Reliabilitätskoeffizienten. Der Wendepunkt liegt etwa bei einem Schwierigkeitsgrad von 0,7, bei CT II und CT III etwas darunter. Das heißt in anderen Worten, bei den Tests, in denen 70% oder weniger der Aufgaben im Klassenschnitt gelöst wurden, sind die Reliabilitätskoeffizienten höher als in den Tests, in denen 70% oder mehr der Aufgaben im Schnitt gelöst wurden.
Am konkreten Beispiel dargestellt haben die KLassen OT 9d und OT 9e mit den numerisch höchsten Schwierigkeitgraden, für die also die Tests sehr leicht waren, sehr niedrige Reliabilitätskoeffizienten und Validitätskoeffizienten, die jeweils kaum mehr akzeptabel sind. Diese Test sollten für diese Klassen ausgesondert werden.
Die bereits in der Literatur beobachtete Tendenz, daß schwierigere Tests größere Validitäts- und Reliabilitätskoeffizienten erbringen, wird in dieser Untersuchung bestätigt.
Es ist jedoch ohne Richtlinien für die Akzeptabilität von Validitäts- bzw. Reliabilitätskoeffizienten nicht möglich, einen 'cut-off-point' festzulegen, von dem aus ein Text als zu leicht zu sehen ist. Die vorliegenden Ergebnisse lassen aber einen solchen Punkt bei ungefähr $\bar{p} = 0,7$ vermuten.

Wenngleich schwierigere Tests eine Zunahme der Validitäts - und Reliabilitätskoeffizienten erbringen, dürfte es nachteilig sein, zu schwierige Tests zu konstruieren. Je weniger Aufgaben in einem CT gelöst werden können, umso geringer wird das textübergreifende Kontextverständnis. Der Proband muß sich immer mehr darauf beschränken, nur Aufgaben mit Hilfe der lokalen Redundanz und der 'transitional probabilities' zu lösen,[26] ohne den Gesamttext zu verstehen, was leicht zu seiner unerwünschten Frustration führt.

---

26 Vgl.: Darstellung der Kontextabhängigkeit von Cloze-items 2.4.4.2.

5.6. Ergänzende Beobachtungen

Außer diesen statistischen Ergebnissen sollen noch einige andere Beobachtungen, die bei der Testdurchführung gemacht wurden, aufgeführt werden.

Gelegentlich wurden von Lehrern bei der Einholung des Lehrerurteils Kommentare zum sprachlichen und schulischen Verhalten von Schülern gegeben. Wegen der Art dieser Arbeit und ihres Umfanges konnten diese nur gelegentlich registriert werden und sollen hier nur dazu dienen, um Problembereiche anzudeuten und zukünftige Forschung darauf hinzuleiten.
In etwa einem Dutzend Fällen (z.B. OT 7e, Schüler Nr. 33,3,11,1; HS HA 8ab-B, Schüler NR. 8; HS HA 7ab-B, Nr. 24, 5 u.a.m.) wurden Schüler von den Lehrern als "Legastheniker" bezeichnet, weshalb ihre schriftlichen Unterrichtsleistungen nicht in die Noten einflossen, und die Schüler deshalb nicht mit 5 oder 6 sondern mit 3 oder 4 oder 5 beurteilt wurden. Man könnte folgern, daß der CT als rein schriftliche Leistungsmessung die Legasthenie erfassen würde. Dies geschah jedoch nur in drei Fällen (OT 7e, Nr. 3; OT 7e, Nr. 11 und HS HA 8ab-B, Nr. 8), in denen der CT gegenüber den Lehrernoten 3, 4 oder 5 als Ergebnis 6 feststellte. In den anderen Fällen entsprach das CT-Ergebnis dem Urteil des Lehrers oder war sogar besser. Die Abhängigkeit der Urteilbildung des CT von der schriftlichen Repräsentation der Sprache kann also aufgrund der Beispiele nicht ohne weiteres nachgewiesen oder widerlegt werden.
Dazu kommt, daß diese Beispiele nicht repräsentativ sind, zumal die Aussagen zur Legasthenie nicht verifiziert sind. Sie lenken aber die zukünftige Forschung auf eine spezielle Problematik bei der Erfassung der Fremdsprachenkompetenz.

Eine erhebliche Diskrepanz zwischen dem Lehrerurteil und dem CT-Ergebnis bestand bei Schüler Nr. 7 der RS A 9f. Der Schüler wurde vom CT mit 'gut', vom Lehrer mit 'mangelhaft' eingestuft. Der Schüler wurde als schüchtern, gehemmt und "sprachlos" beschrieben. Weiter wurde berichtet, der Schüler könnte im Unterricht keinen richtigen Satz bilden. Auf das CT-Ergebnis hin beobachtete der Lehrer den Schüler intensiver und stellte eine von ihm als "Schulangst" bezeichnete Verhaltensstörung fest, die seine Leistungsfähigkeit stark beeinflußte.

Ähnliche Beobachtungen wurden mehrmals gemacht, daß Schüler in CTs gute Ergebnisse lieferten, aber ansonsten kaum ihre Sprachfähigkeit zum Ausdruck brachten. Leider waren diese Fälle zu selten, um als aussagekräftig erfaßt zu werden. Trotzdem kann vermutet werden, daß der CT wegen seiner völligen Unabhängigkeit vom Prüfer und Objektivität unter Umständen in der Lage ist, Kompetenzschichten zu erfassen, die sonst nicht zugänglich sind, weil sie durch psychologische Einflüsse verdeckt sind. Auch dieses Gebiet bedarf weiterer Erforschung.

Ebenso konnten bei der Betrachtung der Antworten bemerkenswerte Beobachtungen gemacht werden. Hier einige Beispiele:

In CT III lauteten die Aufgaben 2 und 3 wie folgt:

"In the evening Tim' s ___2___ said, "Well, Tim, let's go to ___3___ pub."

Mehrmals wurde für 2 die Antwort "friend" gegeben, die für diesen Satz alleine völlig akzeptabel ist. Das Wort "friend" deutet darauf hin, daß der Proband auch den vorhergehenden Satz

"Tim invited his girl friend to his house"

verstanden und in seine Antwort einbezogen hat. Er war jedoch nicht in der Lage, auch den weiterführenden Gedanken des nächsten Satzes

"You're old enough now, and ___4___ want to take my son for ___5___ first drink."

in seine Antwort zu integrieren und zu erkennen, daß Tim ja nicht der Sohn seiner Freundin sein konnte.

Die Antwort anderer Probanden, für 2 "mother" einzusetzen, zeigt, daß zwar der Gedanke des nächsten Satzes einbezogen wurde, aber nicht der übernächste Satz beginnend mit'"But Dad," answered Tim, ...' berücksichtigt wurde. Auch die Antwort "my" für Aufgabe 3 kann zunächst im Rahmen des die nächsten Sätze - sieben an der Zahl - umfassenden Kontextes akzeptiert werden bis zu der Stelle, wo Vater und Sohn darüber diskutieren, in welches Pub sie nun gehen wollen. Es ist allerdings fraglich, inwiefern die falsche Antwort "my" für Aufgabe 3 dem Probanden nun die Möglichkeit für weitere richtige Antworten einschränkt, oder ob dies nicht geschieht. Ob eine solche mehrfache Abhängigkeit zwischen verschiedenen Aufgaben oder nur eine "einfache" Abhängigkeit zwischen einer Aufgabe und dem Kontext besteht, muß von Fall zu Fall entschieden werden.

Dieser Ansatz zur Demonstration textueller Kohärenzen soll nicht weiter verfolgt werden, denn die wesentliche Aussage kann bereits gemacht werden. Der CT ist nicht 'sentence bound', wie Alderson (1978:37) aufgrund statistischer Ergebnisse behauptet (siehe Kapitel 2.4.3., Absatz zur Lückenfrequenz), sondern kontextorientiert.

Ein weiteres Beispiel aus dem gleichen Test CT III zeigt, daß auch die Einbeziehung nicht nur des 'long range' sondern auch des sozio-kulturellen Kontextes zur Lösung des CT notwendig ist:
Aufgabe 12 lautet:

"All right, then Dad. But only _____12_____ drink."

Eine nicht nur syntaktisch und semantisch sondern auch kontextuell völlig akzeptable Antwort wäre hier das Wort "orange". In Kenntnis der sozio-kulturellen Hintergründe muß diese Lösung jedoch abgelehnt werden, denn mit "the first drink" kann nur ein alkoholisches Getränk gemeint sein, zumal sich Tim im letzten Satz nicht als Antialkoholiker erweist.
Man kann selbstverständlich argumentieren, daß die Einbeziehung des Kontextes nicht bei allen Items nötig ist zu ihrer richtigen Beantwortung. Die Frage, wie häufig der Kontext die Akzeptabilität von Antworten bestimme, und wie häufig kontextuell akzeptable Antworten n u r aus der Kenntnis und dem Verstehen des Einzelsatzes heraus möglich waren, wurde daraufhin für alle CTs untersucht. In CT I ist ein Item (Nr. 23), in CT II sind drei Items (Nr. 7, 8 und 23) und in CT III zwei Items (Nr. 7 und 9) immer kontextuell akzeptabel, wenn sie alleine aus der Kenntnis des jeweiligen Satzes beantwortet werden. In diesen Fällen gibt es praktisch nur eine mögliche Antwort.
In allen anderen Fällen könnte man sich für andere Kontexte als die gegebenen auch andere akzeptable Antworten vorstellen.

Neben dem Aufzeigen spezieller Aspekte des CTs, die durch die Statistik nicht erfaßt werden können, werden durch diese Beobachtungen auch Problempunkte der statistischen Betrachtungsweise klar. Insbesondere erhebt sich die Frage, ob ein CT als eine Summe von Einzelaufgaben im traditionellen Sinn gesehen werden kann oder als eine komplexe Gesamtaufgabe mit einer

---

27 Auf Anregung von Prof. Hecht sei hier auf die Möglichkeit verwiesen, diese Art von 'Error Analisis' mit Hilfe der Teilkompetenzen zu systematisieren.

Anzahl von Teilaufgaben. Für diese letztere Annahme spricht die Tatsache der relativ geringen Interdependenz der einzelnen Tests, die allerdings auch anders erklärt werden kann (siehe Kapitel zur Validität). Diese Frage ist jedoch mehr theoretischer Natur, genauer gesagt testtheoretischer Art, und hat keinen Einfluß auf die Aussagekraft der gefundenen Ergebnisse.

6. ZUSAMMENFASSUNG UND FOLGERUNGEN

Der Zweck dieser Studie war es, den CT auf seine Brauchbarkeit hin zu untersuchen. Die Studie wurde nach den Prinzipien der psychometrischen Testtheorie und den fremdsprachentheoretischen Grundlagen durchgeführt. Drei verschiedene Cloze Tests wurden an Schülern der Jahrgangsstufen 5-9 unseres dreigliedrigen Schulsystems auf ihre Gütekriterien hin untersucht. Die Ergebnisse der statistischen Analysen wurden im vorangegangenen Kapitel präsentiert.

Dieses Kapitel enthält eine Zusammenfassung der wichtigen Ergebnisse der statistischen Analysen. Daraus wird geschlossen auf die Annahme, Zurückweisung oder Einschränkung der aufgestellten Forschungshypothesen, und es wird daraus die Beantwortung der eingangs gestellten Forschungshypothesen abgeleitet. Ferner werden die Implikationen der Untersuchungsergebnisse für die zukünftige Test- und Forschungsarbeit dargestellt.

Die beobachteten arithmetischen Mittel entsprachen den erwarteten schulspezifischen Leistungsunterschieden. Sie lagen am höchsten in den Gymnasialklassen, gefolgt von den Realschul- und Hauptschulklassen. Die niedrigsten Werte erzielten jeweils die B-Kurse der Hauptschulen. Die Werte entsprachen ebenso den klassenspezifischen Leistungsstrukturen, wobei in der Regel die höchsten Jahrgangsstufen die höchsten Werte, die unteren die niedrigsten Werte mit den beschriebenen Ausnahmen (5. Klasse HS, 9. Klasse B-Kurs der HS und 9. Klasse RS) erreichten.

Auch die Erwartungen von Reaktionen auf unterschiedliche Textschwierigkeiten wurden bestätigt, indem leichte Texte von den Gruppen besser gelöst, und der schwierige Text mit einem numerisch niedrigeren Testschwie-

rigkeitsgrad bearbeitet wurden. Hinzuweisen ist auf die aufgetauchte Problematik bei der Einschätzung der Textschwierigkeit.
Besonders deutlich wurden die Unterschiede auf den niedrigeren Kompetenzniveaus erfaßt, wogegen sie auf den höheren Kompetenzstufen abflachten.
Die Standardabweichungen wiesen die einzelnen Klassen allgemein als relativ homogene Lerngruppen aus, bestätigten aber für die Gesamtpopulation die Fähigkeit des Tests, zwischen den einzelnen Probanden in hinreichender Weise differenzieren zu können.
Aufgrund dieser Ergebnisse kann für die durchgeführte Untersuchung die Forschungshypothese I, daß der CT in der Lage ist, schul- und jahrgangsspezifische Kompetenzunterschiede in der Fremdsprache aufzuzeigen, angenommen werden.
Die Reliabilitätskoeffizienten der drei verschiedenen CTs, die nach der Split-half Methode und nach der Formel von Kristof berechnet wurden, betrugen:

$r_{tt\ CT\ I}$	0,897
$r_{tt\ KRI\ CTI}$	0,897
$r_{tt\ KRI\ CTII}$	0,876
$r_{tt\ KRI\ CTIII}$	0,898
$r_{tt\ CTII}$	0,878
$r_{tt\ CTIII}$	0,823

Die Werte wurden auch für eine Testlänge von 100 berechnet, der allgemein üblichen Länge zum Vergleich der Präzision verschiedener Tests. Die Werte hierfür betrugen:

$r_{tt\ 100}$ CT I 0,972 CT II 0,967 CT III 0,972

Diese Werte können als sehr gut bezeichnet werden und befriedigen die Ansprüche der Testnormierung bzw. Standardisierung.

Im Bereich der Einzeltests für verschiedene Klassen lagen die Werte niedriger mit gelegentlichen Werten im Negativbereich. Der Großteil der Werte zwischen ca. 0,5 und 0,9 war durchaus akzeptabel für das Niveau informeller Klassentests. Die niedrigen Werte wurden erklärt durch den Störfaktor der irrepräsentablen Variabilität, in anderen Worten durch die unzureichende Größe der S tichprobe und durch die zu geringe Anzahl der Aufgaben.

Auch im Bereich der Einzeltests konnten durch Erhöhung der Aufgabenanzahl bessere Werte erzielt werden. Dies geschah einmal empirisch durch die Zusammenfassung der drei Einzeltests zu 75 Aufgaben, zum anderen durch die mathematische Berechnung der Reliabilität für 100 Aufgaben. Die Ergebnisse, die mit geringen zufallsbedingten Ausnahmen alle über 0,6 lagen, zeigen, daß auch im Bereich der einzelnen Klassentests die Reliabilität genügend hoch sein kann, wenn der Test lang genug ist.

Diese Ergebnisse der Reliabilität sind umso beachtenswerter, als der Test nicht im Sinne der traditionellen Testtheorie manipuliert wurde. Es wurde bei der Testkonstruktion kein Einfluß auf die Güte der Aufgaben ausgeübt. Keine Aufgaben wurden wegen ungenügender Trennschärfe und Schwierigkeit oder wegen möglicherweise bestehender Aufgabeninterkorrelationen eliminiert. Dies hätte, wie bei der Testkonstruktion dargestellt, dem Konstruktionsprinzip des CT widersprochen.

Als Möglichkeit zur Erhöhung der Reliabilität wurde die Methode der Testverlängerung benutzt. Gerade beim CT ist dies durch das Hinzufügen weiterer Textabschnitte mit entsprechender Lückenzahl problemlos und wenig aufwendig. Diese Textabschnitte können entweder eine Fortsetzung des ersten Teils sein, oder aber - wie es hier geschehen ist - können es separate eigene Texte sein, die den Vorteil eines breiteren 'sampling' der Textaufgaben mit sich bringen.

Im Rückblick auf die erzielten Werte, kann auch die Hypothese II, daß der CT ein reliables Meßinstrument ist, unterstützt werden.

Außer durch die dargestellten Meßwerte der Reliabilität kann die Zuverlässigkeit, mit der der CT gleiche Ergebnisse bringt, an den nahezu konstant gleichen Verteilungsmustern der Mittelwerte der verschiedenen Tests in den Einzelklassen abgelesen werden.

Gilt die Reliabilität als erfüllte Voraussetzung für die Validität, so können auch die Ergebnisse der Untersuchung zur Validität als zufriedenstellend gesehen werden. Die Validierung des CT erfolgte an zwei Arten des Lehrerurteils. Von diesen Kriterien ist die Einschätzung der Fremdsprachen-

kompetenz der Schüler gegenüber den Jahresfortgangsnoten als gültiger erkannt worden. Die Validitätskoeffizienten der Einzeltests lagen nicht immer hoch genug, um höhere Ansprüche zu befriedigen. Ursachen können in Leistungs- und Motivationsschwankungen der Schüler oder in einer Vielzahl ungünstiger äußerer Einflüsse bei der Testdurchführung (z.B. Abschreiben, ungünstige Testzeiten, ...) gesehen werden. Gelegentlich wurde auch das Lehrerurteil als mangelhaft erkannt, da z.B. ein Lehrer eine Klasse noch nicht gut genug kannte.

Die Annahme, daß die Validitätskoeffizienten durch eine heterogene Vielfalt von Faktoren beeinflußt werden, wird durch das Fehlen eines konstanten Musters der Werte unterstützt. Die Möglichkeit, daß der CT insgesamt etwas anderes mißt, daß verschiedene Einzeltests wie CT I, II oder III etwas anderes messen als das Kriterium, wird durch die fehlende Konstanz der Validitätskoeffizienten wesentlich verringert.

Durch die Zusammenfassung der Einzeltests werden dieseSchwankungen nicht nur ausgeglichen, sondern die konkurrente Validität insgesamt wesentlich erhöht. Bei der Kombination der drei Tests lagen die Werte um -0,5 oder in der Mehrzahl der Fälle beträchtlich darüber. Ein Ausnahmewert wurde von der Klasse OT 9d mit -0,1 erzielt. Außer der Erklärung durch "Zufall" können zwei andere Möglichkeiten als Ursachen für diesen niedrigen Wert angesehen werden. Einmal können die niedrigen Reliabilitätswerte, die mitbedingt sind durch den 'ceiling effect', die Validität beeinflussen, zum anderen kann auch das Kriterium der Beurteilung der fremdsprachlichen Kompetenz auf dieser relativ höheren Stufe anders orientiert sein. Geht man von einer Taxonomie der Unterrichtsziele zur Erreichung der Fremdsprachenkompetenz aus, deren Stufen nach verschiedenen Jahrgängen ausgerichtet sind, wäre dies eine mögliche Erklärung.

Unter der Voraussetzung, daß das Lehrerurteil als gültiges Kriterium der Fremdsprachenkompetenz anerkannt wird, kann die Forschungshypothese, daß der CT ein gültiges Meßinstrument der Fremdsprachenkompetenz ist - was ausgedrückt wird durch akzeptable Korrelationskoeffizienten der Testbatterie - angenommen werden.

Eine Einschränkung muß für die Durchführung des CT auf höheren Kompetenzstufen gemacht werden. Da die beobachteten Werte nicht als akzeptabel betrachtet wurden, steht eine Validierung hier noch aus.

Sowohl Arbeits- und Zeitaufwand bei der Testerstellung, -durchführung und Auswertung des Tests wurden im Vergleich zu anderen Tests als günstig gesehen. Hypothese IV kann ohne Einschränkungen angenommen werden.

Obwohl die eingangs der Arbeit gestellten Forschungsfragen im Laufe der Ausführungen bereits mehr oder weniger implizit beantwortet wurden, sollen die gefundenen Antworten explizit zusammengestellt werden.

Zur Frage 1:

Ist der CT in unserem Schulsystem bei Schülern mit Deutsch als Muttersprache auch in Anfangsklassen des Englischunterrichts brauchbar?

Die aufgezeigten Werte der Gütekriterien demonstrieren die Brauchbarkeit des CT in den drei Schularten, Hauptschule, Realschule und Gymnasium, besonders für die Anfangsklassen. Schüler mit anderen Muttersprachen als Deutsch wurden von der Untersuchung ausgeschlossen. Die Fremdsprachenkompetenz und ihre Erfassung bei diesen Schülern ist ein Sonderproblem für zukünftige Forschung. Für die praktische Durchführung des Tests ergeben sich aus der Untersuchung folgende Konsequenzen:

Eine Durchführung der gesamten Testbatterie, bzw. mehrerer Einzeltests ist sowohl im Hinblick auf die Testzuverlässigkeit als auf Störfaktoren, die die Validität von Einzeltests beeinflussen können, angeraten. Wegen der einfachen Durchführbarkeit des CT bestehen dabei kaum Probleme. Auf höheren Niveaustufen muß der vorliegende CT mit Einschränkungen betrachtet werden, bzw. es müssen anspruchsvollere Tests entworfen werden. Untersuchungen sind noch nötig bei der Anwendung des CT bei Legasthenikern und verhaltensgestörten Schülern.

Zur Frage 2:

Kann der CT eventuell an Stelle komplizierterer Testmethoden eingesetzt werden, um die Fremdsprachenkompetenz zu messen?

Gültigkeits- und Zuverlässigkeitskoeffizienten sowie seine einfache Durchführbarkeit empfehlen den CT als Alternative zu komplizierteren Testverfahren, selbstverständlich mit den aufgezeigten Einschränkungen. Es sei allerdings darauf hingewiesen, daß für die Schwierigkeiten verhaltensgestörter Kinder sowie von Legasthenikern von keinem der vorhandenen komplizierten Testverfahren Lösungswege angeboten werden.

Zur Frage 3:

Entspricht der Test der Neuorientierung des Englischunterrichts im Hinblick auf das Leitziel der Kommunikativen Kompetenz?

Theoretische Überlegungen sowie die Validierung durch das Lehrerurteil, das ausgerichtet ist auf das allgemein gültige Unterrichtsleitziel der Kommunikativen Kompetenz, zeigen, daß der CT auch diesem Teilziel der Fremdsprachenkompetenz entspricht. Der CT füllt hiermit eine Lücke, denn bislang sind keine Testverfahren daraufhin untersucht und validiert worden.

Zur Frage 4:

Kann der CT zur Zuweisung von Schülern sowohl in verschiedene Schularten als auch in verschiedene Leistungskurse verwendet werden?

Die nachgewiesene Testzuverlässigkeit und Differenzierfähigkeit von Schülern der Gesamtpopulation lassen den CT als geeignetes Instrument zur Zuweisung in verschiedene Schularten und Leistungskurse im Bereich des Englischen erscheinen. Der CT ist eine Bereicherung des Ausleseinstrumentariums, das bislang diese Zwecke erfüllte.

Zur Frage 5:

Könnte der CT dem Lehrer auch im Sinne eines informellen Tests zur Beurteilung der Fremdsprachenkompetenz dienen?

Die Einfachheit des Testverfahrens, insbesondere der Wegfall komplizierter Aufgabenanalyseverfahren und die erreichten Reliabilitäts- und Validitätskoeffizienten und -werte im Bereich von Klassentests lassen den CT als geeignetes Instrument für die Hand des Lehrers im Unterricht erscheinen. Hingewiesen sei auf die Problematik der Textschwierigkeitsbemessung, von der die Güte der erreichten Werte abhängt. Es ist dem Lehrer kaum zuzumuten und für ihn auch kaum durchführbar, einen Test vor der Anwendung auf die Angemessenheit des Schwierigkeitsgrades hin zu erproben. Andererseits ist von einem erfahrenen Lehrer zu erwarten, daß er den für seine Klasse geeignetsten Schwierigkeitsgrad subjektiv erfassen kann.

Zur Frage 7:

Was ist das beste Format des CT?

Die Frage soll spezifiziert werden für das untersuchte Standardverfahren

des CT. Aufgrund der untersuchten Variablen kann gesagt werden, daß der CT die besten Ergebnisse bei einer Durchführung mehrerer Einzeltests mit ca. 25 Aufgaben von mittlerer bis größerer Textschwierigkeit liefert. Ausgenommen sind die 9. Gymnasialklassen, für die die Tests zu leicht sind. Untersuchungen für andere Verfahren mit anderen Variablen, z.B. kleinere oder größere Lückenfrequenz oder besondere Methoden zur Lückenbildung, stehen noch aus.

Zur Frage 6:

Was sind Vor- bzw. Nachteile des CT?

Diese Frage ist an die letzte Stelle gestellt, da aus ihrer Beantwortung zugleich weitere Ausblicke auf zukünftige Forschungsmöglichkeiten gelenkt werden können.

Wie demonstriert wurde, sind die Hauptvorteile des CT seine hohen Werte der Gütekriterien. Daraus darf aber nicht seine Alleingültigkeit abgeleitet werden.

Bislang bietet der CT keine ausgearbeiteten Verfahren zur Diagnostizierung von Schwächen der Fremdsprachenkompetenz. Eine sorgfältige 'Error Analysis' könnte eventuell dorthin führen. Dieses Gebiet muß zunächst weiterhin den traditionellen Testverfahren überlassen bleiben. Außerdem sind die theoretischen und empirischen Grundlagen nicht weit genug fortgeschritten, um daraus mit Sicherheit die Meßfähigkeit sämtlicher Teile der Fremdsprachenkompetenz in a l l e n Bereichen, insbesondere im mündlichen Bereich, profund genug nachzuweisen. Wie eingangs erwähnt und festgelegt, erfolgte die Untersuchung nur im schriftlichen Bereich.

Zukünftige Forschung muß deshalb ausgerichtet sein auf die

E r f o r s c h u n g d i a g n o s t i s c h e r M ö g l i c h k e i t e n des CT,

E r f a s s u n g s m ö g l i c h k e i t e n auch des m ü n d l i c h e n B e r e i c h s der Fremdsprachenkompetenz,

E r f a s s u n g h ö h e r e r K o m p e t e n z n i v e a u s , und

P r o b l e m a t i k bestimmter S c h ü l e r g r u p p e n wie Legastheniker, verhaltensgestörter Schüler und Schüler mit anderen Muttersprachen als Deutsch, die aber Englisch als Fremdsprache in unserem Schulsystem lernen.

7. APPENDIX

7.1. Testbeispiele mit Lösungsblättern

Name _____ 183 Wieviele Monate Englischunterricht? _____
(Familienname, Vorname)

Geburtsdatum _____ Alter _____ Schule _____

Muttersprache _____ Klasse _____

Zeit im englischsprachigen Ausland _____ Letzte Zeugnisnote in Englisch _____

Letzte Zeugnisnote in Deutsch _____

Welche Englischnote würdest Du Dir selber geben? _____

Noten der letzten Englischtests: Art der Arbeit _____ Note _____

Art der Arbeit _____ Note _____ Art der Arbeit _____ Note _____

(Art der Arbeit z. B. Schulaufgabe, Probe, Ex.../Diktat, Aufsatz, Übersetzung, Vermischtes...).

Anleitung:

1)Lies den Text und versuche, den Inhalt zu verstehen.

2)Trage in jeden freien Raum am rechten Rand das Wort ein,von dem Du glaubst,daß es am besten in die Lücke paßt.Achtung: NUR EIN WORT!Zusammengezogene Wörter(don't/John's) zählen als ein Wort!

3)Rechtschreibefehler zählen nicht gegen Dich .

4)Überprüfe Deine Antworten.

Beispiel:David is eleven years old. His sister Shirley is eighteen years (Beisp.) and his brother John twenty-two.

(Beisp.) old

Who is who?

Ruth Richmond and John are standing at the window of their classroom. "Who is that girl down there?" John 1 .
"Oh, that is Grace," Ruth tells 2 , "she is the new girl." Robert 3 over John's shoulder and says, "Her 4 isn't Grace. It's Barbara."
"Her name 5 Grace Carter, Robert. Her Mother calls 6 Grace."
"No, her Mother calls her 7 ."
"Look, she is coming up," John 8 .
The new girl enters the classroom. 9 has blond hair, brown eyes, and 10 is wearing a green dress.
"Hello, 11 ," Ruth says.
"I'm not Grace. I'm 12 ."
"You see, Ruth," Robert says, "that's 13 I say, her name is Barbara."
14 , John and Robert are going down 15 the play - ground. There is the new 16 again.
"Come along, Barbara," Ruth says.
" 17 not Barbara. My name is Grace."
"Well, I say! John says, "this girl 18 two names. All right, Grace and Barbara, let's 19 up to the classroom."
The 20 enter the classroom. Robert stops. His 21 and mouth are wide open. There 22 a girl in the classroom with 23 same eyes, the same hair, the 24 dress as Grace. Robert looks at 25 two girls. He puts his hands over his head. "Ha,ha," John laughs, "they are twins. Grace and Barbara are twins!"

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____
13. _____
14. _____
15. _____
16. _____
17. _____
18. _____
19. _____
20. _____
21. _____
22. _____
23. _____
24. _____
25. _____

"Who is Who?"

1. asks	says
2. him	
3. looks	
4. name	
5. is	
6. her	
7. Barbara	
8. says	
9. she	
10. she	
11. Grace	
12. Barbara	
13. what!	Barbara right
14. Ruth, him, was	
15. to	
16. girl	pupil schoolgirl
17. don	no
18. has	
19. go	run walk
20. children [illegible]	boys group friends pupils
21. eyes	
22. is	
23. the	
24. same	
25. the	

Name _____ Wieviele Monate Englischunterricht? _____
(Familienname, Vorname)

Geburtsdatum _____ Alter _____ Schule _____

Muttersprache _____ Klasse _____

Zeit im englischsprachigen Ausland _____ Letzte Zeugnisnote in Englisch _____

Letzte Zeugnisnote in Deutsch _____

Welche Englischnote würdest Du Dir selber geben? _____

Noten der letzten Englischtests: Art der Arbeit _____ Note _____

Art der Arbeit _____ Note _____ Art der Arbeit _____ Note _____

(Art der Arbeit z. B. Schulaufgabe, Probe, Ex.../Diktat, Aufsatz, Übersetzung, Vermischtes...).

Anleitung:

1)Lies den Text und versuche, den Inhalt zu verstehen.

2)Trage in jeden freien Raum am rechten Rand das Wort ein,von dem Du glaubst,daß es am besten in d
Lücke paßt.Achtung: NUR EIN WORT!Zusammengezogene Wörter(don't/John's) zählen als ein Wort!

3)Rechtschreibefehler zählen nicht gegen Dich .

4)Überprüfe Deine Antworten.

Beispiel:David is eleven years old. His sister Shirley is eighteen years (Beisp.) and his brother John twenty-two.

(Beisp.)

Pat at the Grocer's

Today is Saturday and Pat has no school. So Mrs Black says to her (1) :"I have a lot to do, Pat, (2) can help me to make (3) apple-pie, but we have no flour (4) no sugar in the house. Please, (5) to the grocer's and get me (6) pounds of flour, one pound (7) sugar, two pounds of apples, half (8) pound of butter and a dozen (9) ."

Pat gets her mother's shopping-bag and (10) :" "I have no money, Mummy." Mrs Black (11) the door of her cupboard, takes (12) a pound note and hands it (13) Pat.

Pat walks up the road (14) Mr Green's shop. At the corner of (15) road she suddenly stops. "Let me (16) , -what have I to get?- One (17) of apples -or two? Two pounds of flour (18) one?Two, I think.One pound (19) sugar - and what else?

A pound of butter? (20) is something else! I must fetch (21) for the pie."

When Pat gets (22) the grocer's ,Mr Green asks her :"(23) can I do for you, Pat?"- "(24) mother wants to make an apple-pie, so (25) give me two pounds of butter, half a pound of apples, and one pound of sugar."

"Are you sure, Pat?Is it not:two pounds of apples, half a pound of butter and don't forget two pounds of flour and a dozen eggs."-"Oh, thank you, Mr Green, that's it.

"Isn't Mr Green a nice and helpful man " she thinks and runs home.

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____
13. _____
14. _____
15. _____
16. _____
17. _____
18. _____
19. _____
20. _____
21. _____
22. _____
23. _____
24. _____
25. _____

"Pat at the Grocer's"

1. daughter
2. you
3. an
4. and
5. go
6. two
7. of
8. a
9. eggs
10. says
11. opens
12. out
13. to
14. to
15. the
16. see think remember
17. pound
18. or
19. of
20. there
21. eggs something apples
22. to
23. what
24. [illegible]
25. bears

Name _____ 187 _____ Wieviele ~~Jahre~~ Monate Englischunterricht? _____
(Familienname, Vorname)

Geburtsdatum _____ Alter _____ Schule _____

Muttersprache _____ Klasse _____

Zeit im englischsprachigen Ausland _____ Letzte Zeugnisnote in Englisch _____

Letzte Zeugnisnote in Deutsch _____

Welche Englischnote würdest Du Dir selber geben? _____

Noten der letzten Englischtests: Art der Arbeit _____ Note _____

Art der Arbeit _____ Note _____ Art der Arbeit _____ Note _____

(Art der Arbeit z. B. Schulaufgabe, Probe, Ex.../Diktat, Aufsatz, Übersetzung, Vermischtes...).

Anleitung:

1)Lies den Text und versuche, den Inhalt zu verstehen.

2)Trage in jeden freien Raum am rechten Rand das Wort ein, von dem Du glaubst, daß es am besten in d
Lücke paßt. Achtung: NUR EIN WORT! Zusammengezogene Wörter (don't/John's) zählen als ein Wort!

3)Rechtschreibefehler zählen nicht gegen Dich .

4)Überprüfe Deine Antworten.

Beispiel: David is eleven years old. His sister Shirley is eighteen years (Beisp.) and his brother John twenty-two.

(

Tim's First Drink

It was Tim Price's eighteenth birthday last week. Tim invited his girl friend Janet __1__ his house. 1. _____

In the evening Tim's __2__ said,"Well, Tim, let's go to __3__ pub. You're old enough now, and __4__ want to take my son for __5__ first drink. Janet can't come - she's still __6__ young - but your mother will stay __7__ home with her." 2. _____ 3. _____ 4. _____ 5. _____ 6. _____ 7. _____

"But, Dad," answered __8__, "Janet and I were just going __9__ go out. The friends we want to meet are already waiting for us." 8. _____ 9. _____

"Oh, Tim! __10__ was looking forward to a drink __11__ you." 10. _____ 11. _____

"All right, then, Dad. But only __12__ drink." 12. _____

Father and son left the __13__. "Which pub are we going to?" __14__ Tim. "The pub I've chosen for __15__ evening is The Black Pig. I always __16__ to The Ship of Gold myself, __17__ there are lots of young people __18__ The Black Pig." 13. _____ 14. _____ 15. _____ 16. _____ 17. _____ 18. _____

It will be __19__ for you there." 19. _____

"But, Dad, I'd __20__ to go to The Ship of Gold." 20. _____

"No, son, __21__ sure you'll like it better at The Black Pig." 21. _____

Tim and __22__ father entered The Black Pig. A __23__ of young people at the bar __24__ round. 22. _____ 23. _____ 24. _____

"Hey, Tim, you're late!" "Happy __25__ ,Tim!" "Where's Janet?" 25. _____

The barman smiled. "What will it be this time,Tim? A pint of Guinness, as usual?

Tim's First Drink

1. [illegible]	
2. [illegible]	dad
3. [illegible]	a
4. I	
5. [illegible]	
6. [illegible]	
7. [illegible]	
8. [illegible]	
9. [illegible]	
10. [illegible]	
11. [illegible]	
12. one	
13. house	flat apartement
14. called	said
15. this	the
16. [illegible]	
17. but	
18. at	in
19. nice	nice better good interesting
20. like	
21. [illegible]	
22. he	
23. crowd [illegible]	group
24. turned	looked
25. birthday	

7.2. Gesamtübersicht der statistischen Ergebnisse der Durchführung einer Testbatterie

bei Schülern der Jahrgangsstufe 5 bis 9 aller Schularten

CT I II III

N	$\bar{X}_{CT\,123}$ RANGE	SD	$r_{CT1/CT2}$	$r_{CT1/CT3}$	$r_{CT2/CT3}$	$r_{CT123/E}$	$r_{CT123/TE}$	r_{tt} / $r_{tt\,KRI}$	$\bar{p}$
14	8,9 4,7-11,3	2,196	0,365	0,168	0,515	-0,304	-0,798	0,771 0,799	0,36
29	14,26 9,6-17,6	2,004	0,244	0,107	0,416	0,361	-0,623	0,828 0,833	0,57
26	16,34 7-22,3	3,311	0,693	0,558	0,417	-0,599	-0,909	0,812 0,826	0,65
16	8,59 4,6-14,3	2,771	0,699	0,584	0,656	-0,579	-0,847	0,717 0,754	0,34
26	19,78 15,3-28	1,880	0,532	0,329	0,540	-0,343	-0,788	0,610 0,658	0,79
32	20,84 17,3-24	1,684	0,307	0,053	0,408	-0,764	-0,946	0,606 0,631	0,83
10	12,64 10-16,6	2,075	0,048	0,455	0,58	-0,565	-0,812	0,598 0,681	0,51
19	10,73 4-19,3	3,643	0,555	0,508	0,317	-0,543	-0,658	0,734 0,764	0,43
28	17,08 13,6-22,4	2,144	0,443	0,504	0,416	-0,537	-0,760	0,793 0,797	0,68
30	20,63 12-24,8	5,068	0,647	0,618	0,430	-0,549	-0,887	0,602 0,630	0,82
12	13,858 12-18,6	1,786	0,122	0,493	0,298	-0,579	-0,595	0,342 0,441	0,55
18	18,272 14-29,6	1,756	0,607	0,260	0,547	-0,319	-0,767	0,170 0,268	0,73
18	14,578 11,6-19,4	1,908	-0,163	-0,085	0,517	-0,127	-0,946	0,647 0,596	0,58
15	19,187 14-22	2,399	0,415	0,221	0,413	-0,617	-0,743	0,784 0,781	0,77
28	21,268 17,6-24,1	1,79	0,138	0,124	0,591	-0,519	-0,490	0,811 0,816	0,85
21	13,738 3,3-20,4	4,213	0,744	0,608	0,775	-0,486	-0,804	0,861 0,874	0,55
10	6,11 4,3-9,2	1,980	0,289	0,013	-0,323	-0,343	-0,978	0,748 0,805	0,24
10	15,05 9,6-20,6	4,659	0,743	0,814	0,825	-0,606	-0,933	0,876 0,903	0,60
18	19,93 17-22,6	2,198	0,399	0,698	0,305	-0,62	-0,73	0,554 0,604	0,79
31	23,27 19,4-24,7	0,990	0,352	0,006	0,293	-0,157	-0,117	0,438 0,445	0,93
29	22,45 20,3-24,7	1,159	0,240	0,002	0,071	0,064	-0,521	0,476 0,506	0,90

ERKLÄRUNGEN

HS 5	Hauptschule, 5. Schülerjahrgang
RS 7	Realschule, 7 Schülerjahrgang
GY 9	Gymnasium, 9. Schülerjahrgang
N	Anzahl der Probanden
T_S	Testsequenz der einzelnen Clozetests innerhalb der Batterie
Z	Zeitpunkt der Durchführung
$\bar{p}$	mittlerer Schwierigkeitsgrad, Berechnung nach Formel (2)
$\bar{X}_{CT}$	arithmetisches Mittel der Clozetestrohwerte Berechnung nach Formel (1)
RANGE	Angabe der Variationsweite vom kleinsten empirischen Meßwert zum größten empirischen Meßwert X_{min} — X_{max}
SD	Standardabweichung der Clozetestrohwerte Berechnung nach Formel (3)
$\bar{X}_E$	arithmetisches Mittel der Schulnoten (Jahresfortgangsnoten) Berechnung nach Formel (1)
$\bar{X}_{TE}$	arithmetisches Mittel der Lehrerbeurteilung Berechnung nach Formel (1)
$r_{CT/E}$ CT/TE	Korrelation zw. Clozetestrohwerten und Jahresfortgangsnoten bzw. Lehrerbeurteilung Berechnung nach Formel (5)
r_{tt}	Reliabilitätskoeffizient, Berechnung nach Formel (6) mit Aufwertung nach Spearman-Brown
r_{ttKRI}	Reliabilitätskoeffizient, Berechnung nach Formel (7)

FORMEL (6)

Reliabilitätskoeffizient-r_{tt} (Split-Half)

$$r_{tt} = \frac{2 \cdot r_{12}}{1 + r_{12}}$$

wobei: r_{12} Korrelation zwischen gerader und ungerader Testhälfte

FORMEL (7)

Reliabilitätskoeffizient-$r_{tt\,KRI}$ (nach Kristof)

$$r_{tt\,KRI} = \frac{2}{N-1} + \frac{N-3}{N-1} \cdot \frac{4 s_1 s_2 r_{12}}{s_1^2 s_2^2 + 2 s_1 s_2 r_{12}}$$

n Anzahl der Probanden

$s_1 s_2$ Standardabweichungen der geraden und ungeraden Testhälften

FORMELN

FORMEL (1):

Mittelwert-$\bar{X}$ (arithmetisches Mittel) der Rohwerte X

$$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

FORMEL (2)

mittlerer Schwierigkeitsindex-$\bar{p}$

$$\bar{p} = \frac{\sum p}{n}$$

n Zahl der Items

p Schwierigkeitsindex-$100 \cdot \frac{N_R}{N}$

N_R-Anzahl der Probanden, die die Aufgabe richtig beantwortet haben

N Gesamtzahl der Probanden

FORMEL (3)

Standardabweichung-SD mit n-1 Gewichtung

$$SD = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n^2}\right) \cdot \frac{n}{n-1}}$$

n-Anzahl der Meßwerte

x_i i-ter Meßwert der Häufigkeitsverteilung

FORMEL (4)

Varianz-Var mit n Gewichtung

$$Var = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n^2}\right)^2$$

FORMEL (5)

Korrelationskoeffizient-r (Pearson-Produkt-Moment-Korrelation)

$$r = \frac{msx}{sy}$$

wobei: sx^2, sy^2 = Varianz der x, y Werte

$$m = \frac{\frac{\sum x_i \sum y_i}{n} - \sum x_i y_i}{\frac{\left(\sum x_i\right)^2}{n} - \sum x_i^2}$$

I WHO IS WHO

Schulart / Jahrgang	Schule / Klasse	N	T_s	Z	$\bar{P}$	$\bar{X}_{CT}$ / RANGE	SD	$\bar{X}_e$ / $\bar{X}_{TE}$	SD_e / SD_{TE}	$r_{CT/E}$ / $r_{CT/TE}$	r_{tt} / $r_{tt\,KR}$
HS 5	LO 5a	14	1	[illegible]	[illegible]	12,5 / [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GY 5	BB 5c	29	1	3/79	0,61	13,48 / 10-21	2,814	2,73 / 3,27	0,96 / 0,8	[illegible]	[illegible]
GY 5	OT 5b	26	1	9/79	0,70	17,62 / 7-24	4,78	2,88 / 3,79	0,99 / 1,01	[illegible]	[illegible]
HS 6	FFB 6e AKurs	16	1	5/79	0,41	10,313 / 6-18	3,98	3,5 / 3,28	0,52 / 1,14	[illegible]	[illegible]
GY 6	BB 6b	26	1	3/79	0,82	20,54 / 16-24	2,063	3,38 / 3,40	0,69 / 0,68	[illegible]	[illegible]
GY 6	OT 6e	32	1	7/79	0,86	21,50 / 14-25	2,436	3,25 / 3,52	0,916 / 1,004	[illegible]	[illegible]
HS 7	FFB 7d AKurs	10	1	[illegible]	[illegible]	[illegible] / 4-17	2,514	3,10 / 3,40	0,87 / 1,14	[illegible]	[illegible]
HS 7	HA 7ab BKurs	19	1	5/79	[illegible]	[illegible]	[illegible]	[illegible] / 1,87	[illegible]	[illegible]	[illegible]
RS 7	A 7a	28	2	4/79	0,78	[illegible] / 3-23	[illegible]	3,78 / 3,57	[illegible] / 0,71	[illegible]	0,688 / [illegible]
GY 7	OT 7e	30	1	7/79	0,85	21,37 / 8-25	3,429	[illegible] / 3,42	1,11 / 1,23	[illegible]	0,738 / 0,755
HS 8	FFB 8d AKurs	12	1	2/79	0,57	14,17 / 10-19	[illegible]	3,08 / [illegible]	[illegible] / 0,81	[illegible]	[illegible] / 0,443
HS 8	HA 8ab AKurs	18	1	7/79	0,75	18,67 / 15-22	2,024	2,83 / 3,25	0,92 / [illegible]	[illegible]	0,556 / 0,605
HS 8	HA 8ab BKurs	18	1	7/79	0,53	13,17 / [illegible]	[illegible]	2,94 / [illegible]	0,54 / 0,96	[illegible]	[illegible]
RS 8	A 8a	[illegible]	2	[illegible]	[illegible]	22,2 / 18-24	[illegible]	3,27 / 3,3	[illegible] / 0,77	[illegible]	[illegible]
GY 8	OT 8f	28	3	9/79	[illegible]	22,82 / 18-25	[illegible]	3,21 / 3,30	0,96 / 0,92	[illegible]	[illegible] / 0,722
HS 9	LO 9ab AKurs	21	1	12/79	0,58	14,48 / 4-20	5,173	2,86 / 3,19	0,72 / 0,79	[illegible]	[illegible] / 0,907
HS 9	LO 9ab BKurs	[illegible]	1	10/79	[illegible]	[illegible] / 3-16	3,836	4,6 / 6,3	0,51 / 0,86	[illegible]	[illegible] / 0,865
HS 9	FFB 9 AKurs	10	1	10/79	0,66	[illegible] / 8-22	[illegible]	[illegible] / 3,45	[illegible] / 0,38	[illegible]	[illegible] / 0,447
RS 9	A 9f	18	1	9/80	0,85	[illegible] / 12-25	1,936	3,78 / 3,83	0,65 / [illegible]	0,692 / [illegible]	0,680 / 0,714
GY 9	OT 9d	31	1	7/80	0,95	23,78 / 19-25	1,175	3,22 / 3,27	0,884 / [illegible]	[illegible]	0,684 / 0,674
GY 9	OT 9e	29	1	7/80	0,91	21,97 / 18-25	1,592	3,41 / 3,51	[illegible] / 1,128	[illegible]	0,414 / [illegible]

II PAT AT THE GROCER'S

Schulart / Jahrgang	Schule / Klasse	N	T_s	Z	$\bar{P}$	$\bar{X}_{CT}$ / RANGE	SD	$\bar{X}_e$ / $\bar{X}_{TE}$	SD_e / SD_{TE}	$r_{CT/E}$ / $r_{CT/TE}$	r_{tt} / $r_{tt\,KR}$
HS 5	[illegible]	4	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GY 5	[illegible]	[illegible]	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GY 5	[illegible]	26	2	9/79	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
HS 6	[illegible]	16	2	[illegible]	[illegible]	[illegible]	3,556	3,6 / 3,28	0,6 / 1,14	[illegible]	[illegible]
GY 6	[illegible]	[illegible]	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GY 6	[illegible]	[illegible]	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0,597	0,806
HS 7	[illegible]	[illegible]	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0,324 / -0,455	0,592 / 0,682
HS 7	[illegible]	[illegible]	2	5/79	0,43	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
RS 7	[illegible]	28	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GY 7	[illegible]	30	2	[illegible]	[illegible]	[illegible] / 13-25	[illegible]	3,27 / [illegible]	1,11 / [illegible]	[illegible] / 0,285	[illegible] / 0,428
HS 8	[illegible]	12	2	2/79	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
HS 8	[illegible]	18	2	7/79	[illegible]	[illegible]	2,312	[illegible]	[illegible]	[illegible]	[illegible]
HS 8	[illegible]	18	2	7/79	[illegible]	15,94	2,182	3,94 / [illegible]	[illegible]	[illegible]	[illegible]
RS 8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3,27 / [illegible]	[illegible]	[illegible]	[illegible]
GY 8	[illegible]	[illegible]	2	[illegible]	[illegible]	[illegible]	[illegible]	3,21 / 3,30	0,96 / 0,92	[illegible]	[illegible]
HS 9	[illegible]	21	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
HS 9	[illegible]	[illegible]	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
HS 9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0,944	0,823
RS 9	[illegible]	18	2	9/80	[illegible]	[illegible]	[illegible]	3,78 / [illegible]	0,65 / [illegible]	[illegible]	[illegible]
GY 9	[illegible]	[illegible]	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0,884 / [illegible]	[illegible]	[illegible]
GY 9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0,298	0,440

III TIM'S FIRST DRINK

Schulart / Jahrgang	Schule / Klasse	N	T_s	Z	$\bar{P}$	$\bar{X}_{CT}$ / RANGE	SD	$\bar{X}_e$ / $\bar{X}_{TE}$	SD_e / SD_{TE}	$r_{CT/E}$ / $r_{CT/TE}$	r_{tt} / $r_{tt\,KR}$
HS 5	[illegible]	4	3	7/80	[illegible]	7,14 / [illegible]	1,875	3,5 / 3,53	0,65 / 0,86	0,063 / 0,572	0,401 / 0,437
GY 5	BB 5c	29	2	7/79	0,37	9,21 / 4-17	3,589	2,93 / 3,27	0,96 / 0,8	0,089 / 0,325	0,494 / 0,504
GY 5	OT 5b	26	3	9/79	0,58	14,46 / 6-23	3,679	2,88 / 3,29	0,99 / 1,01	0,280 / -0,628	0,767 / 0,704
HS 6	FFB 6e AKurs	16	3	6/79	0,21	5,44 / 3-9	1,861	3,5 / 3,28	0,5 / 1,14	0,104 / -0,565	0,782 / -0,428
GY 6	BB 6b	26	3	4/79	0,74	18,58 / 14-23	2,626	3,38 / 3,40	0,69 / 0,68	-0,519 / 0,798	0,428 / 0,466
GY 6	OT 6e	32	3	7/79	0,82	20,80 / 17-24	2,016	3,28 / 3,52	0,916 / 1,004	-0,454 / 0,634	0,843 / 0,374
HS 7	FFB 7d AKurs	10	3	12/79	0,45	11,2 / 7-16	2,70	3,10 / 3,40	0,87 / 1,19	-0,479 / -0,612	0,362 / 0,448
HS 7	HA 7ab BKurs	19	3	7/79	0,39	9,79 / 3-2[illegible]	4,364	3,58 / 3,87	0,84 / 0,83	-0,361 / 0,231	0,758 / 0,782
RS 7	A 7a	28	[illegible]	4/79	0,45	11,32 / 8-20	2,763	3,18 / 3,59	0,61 / 0,91	-0,583 / 0,666	0,458 / 0,498
GY 7	OT 7e	30	3	7/79	0,79	19,93 / 15-24	2,243	3,27 / 3,42	1,11 / 1,23	-0,352 / 0,638	0,511 / 0,544
HS 8	FFB 8d AKurs	12	3	7/79	0,47	[illegible] / 9-18	2,552	[illegible] / 3,33	0,51 / 0,81	-0,542 / -0,478	0,593 / [illegible]
HS 8	HA 8ab AKurs	18	3	7/79	0,68	[illegible] / [illegible]	2,449	2,83 / 3,25	0,92 / 0,91	0,052 / -0,566	0,264 / 0,546
HS 8	HA 8ab BKurs	[illegible]	3	7/79	[illegible]	[illegible] / [illegible]	[illegible]	3,94 / 3,6	0,54 / 0,96	0,235 / 0,864	0,670 / [illegible]
RS 8	A 8a	15	1	1/79	0,65	16,13 / 8-22	3,944	3,27 / 3,3	0,70 / 0,77	-0,606 / -0,762	0,824 / 0,841
GY 8	OT 8f	[illegible]	1	4/79	[illegible]	[illegible] / 12-23	2,855	3,2[illegible] / 3,30	[illegible] / 0,92	0,48 / 0,482	0,70 / 0,709
HS 9	LO 9ab AKurs	[illegible]	2	[illegible]	0,47	[illegible] / [illegible]	4,857	2,86 / 3,19	0,72 / 0,79	0,279 / 0,742	[illegible] / 0,865
HS 9	LO 9ab BKurs	[illegible]	3	2/80	[illegible]	[illegible] / [illegible]	[illegible]	4,6 / 5,3	0,51 / 0,86	0,380 / 0,150	[illegible] / 0,009
HS 9	FFB 9 AKurs	[illegible]	3	1/80	0,57	[illegible] / [illegible]	[illegible]	[illegible] / 3,45	0,73 / 0,38	-0,623 / -0,916	0,876 / 0,901
RS 9	A 9f	18	3	9/80	0,72	[illegible] / [illegible]	3,298	3,78 / 3,83	0,65 / 1,03	0,490 / 0,673	0,451 / 0,515
GY 9	OT 9d	31	3	7/80	0,89	22,32 / 19-25	1,301	3,22 / 3,27	0,884 / 0,93	-0,094 / -0,131	0,367 / 0,293
GY 9	OT 9e	[illegible]	2	7/80	0,84	[illegible] / 8-25	[illegible]	[illegible] / 3,51	0,725 / 1,29	[illegible] / 0,336	[illegible] / 0,299

LITERATURVERZEICHNIS

Aitken, K.G., "Problems in cloze testing reexamined", TESL Reporter, 1975, 8 (2), 9; 16-17.

Aitken, Kenneth, G., "Using Cloze Procedure as an Overall Language Proficiency Test", TESOL Quarterly, Vol. 11, Nov./March 1977, 59-67.

Alderson, J.C., A study of the cloze procedure with native and non-native speakers of English, (Edinburgh, 1978).
Unpublished Ph. D. Thesis

Alderson, J.C., "The Cloze Procedure and Proficiency in English as a Foreign Language", TESOL, Vol. 13, June 1979, 219-227.

Alderson, J.C., "Native and Nonnative Speaker Performance on Cloze Tests", Language Learning, Vol. 30, Nov. / June 1980, 59-76.

Allen, V.F., "Toward a thumb-nail test of English competence", TESOL Quarterly, 1968, 2 (4), 240-242.

Allen, H.B., &Campbell, R.N., Teaching English as a second language: a book of readings (New York, ^{2}1972)

Allen, J.P.B., & Alan Davies, Testing and experimental methods, (London, 1977), The Edinburgh Course in Applied Linguistics, Vol. 4.

Anderson, Jonathan, Application of cloze procedure to English learned as a foreign language. Unpublished doctoral dissertation, University of New England, Australia, 1969.

Anderson, Jonathan,"A psycholinguistic approach to the measurement of reading comprehension materials".Kivung, Vol. 3 (2), 83-94.

Anderson, Jonathan, "A Technique for Measuring Reading Comprehension and Readability", ELT, Vol. 25 (2), 1971, 178-182.

Anderson, Jonathan, "The application of cloze procedure to English learned as a foreign language in Papua and New Guinea", ELT, Vol. XXVII (1), 66-72.

Anderson, J., Psycholinguistic experiments in foreign language testing, St. Lucia, Queensland: University of Queensland Press, 1976.

Anshen, Frank, Statistics for linguists.(Rowley, Mass., 1978)

Arndt, Horst,"Zur Problematik der Leistungskontrolle im fremdsprachlichen Unterricht", Der fremdsprachliche Unterricht, Heft 14, Mai 1970:, 2-15.

Austin, J.L., How to do things with words. (London, 1962)

Bailey, W., "Practical application of Cloze Tests", Zielsprache Englisch, Heft 4,(München, 1976), 24-28, und 135-145.

Belser, "Wie beurteilt man Schultests?", Ingenkamp, Tests in der Schulpraxis, (Weinheim, ^{3}1973).

Bormuth, J.R. (1964), "Experimental application of Cloze Tests", in IRA Conference, Proc. 9, 303-306.

Boyce, Max (1976), "A comprehensive Bibliography of the Cloze Procedure", Part B, ERIC, Arlington, Va., ED 127 580, Sept. 1976.

Breuer, V. (1974), Englisch Tests und objektive Leistungsmessung Dornburg-Frickhofen.

Brière, E.J. (1972), "Are we really measuring proficiency with our foreign language tests?", in Teaching English as a second language: a book of readings, ed. H.B. Allen & R.N. Campbell, New York, 321-330.

Brière, E.J. (1969), "Current trends in second language Testing", TESOL Quarterly, 3 (4), 333-340.

Brière, E.J. & Hinofotis, Frances (1979), Concepts in language testing: some recent studies , TESOL. Washington, D.C.

Brière / Clausing / Senko / Parcell (1978), "A Look at Cloze Testing Across Languages and Levels", MLJ, Jan./Febr., 1978, 23-26

Brings, F. (1972), "Der Leistungstest. Ein Literaturbericht." Praxis des neusprachlichen Unterrichts. Heft 1/1972.

Brown, F.L. & Amos, J.R. & Mink, O.G. (^{2}1975), Statistical Concepts: A Basic Program. New York, San Francisco, London.

Brown, J.P. (1978), Correlational Study of Four Methods For Scoring Cloze Tests. Unpublished Master's Thesis. University of California, Los Angeles, 1978.

Brown, J.D. (1979), "Correlational Study ofFour Methods for Scoring Cloze Tests", Paper presented at the TESOL CONVENTION, 1979 in Boston.

Brumfit, C.J & Johnson, K. (1979), The Communicative Approach to Language Teaching (Oxford, 1979)

Campbell, D.T. & Fiske, D.W. (1959), "Convergent and discriminant validation by the multitrait-multimethod matrix", Psychological Bulletin 56 (2):81-105.

Carpenter, P.A. & Just, M.A. (1975), "Sentence Comprehension: A Psycholinguistic Model of Verification", Psychological Review, Bd. 82:45-73.

Carroll, J.B. & Carton, A.S. & Wilds, C.P. (1959), "An investigation of cloze items in the measurement of achievement in foreign languages", College Entrance Examination Board Research and Development Report. Laboratory for Research in Instruction, Harvard University.

Carroll, J.B. (1961), "Fundamental consideration in testing for English language proficiency of foreign students testing" H.B. Allen & R.W. Campbell (1972), Testing the English proficiency of foreign students (Washington, 1972), 313-320.

Carroll, J.B. (^{2}1970), "The Psychology of Language Testing", A. Davies, Language Testing Symposium: A Psycholinguistic Approach. London, ^{2}1970:46-70

Carroll, J.B. (1972), "Defining language comprehension: some speculations", R.O. Freedle & J.B. Carroll, Language comprehension and the acquisition of knowledge (Washington, D.C., 1972), 1-29.

Carroll, J.B. (1973), "Implications of Aptitude Test Research and Psycholinguistic Theory for Foreign Language Teaching". Linguistics 112, 5-14.

Chauncey, H. & Dobbin, J.E. (1968), Der Test im modernen Bildungswesen, Stuttgart, 1968.

Chihara, T. & Oller, J. & Weaver, K. & Chavez-Oller, M.A. (1977), "Are Cloze Items Sensitive To Constraints across Sentences?", Language Learning, Vol. 27, Nov., 63-73.

Chomsky, N. (1973), "Linguistic Theory", J.W. Oller & J.C. Richards, Focus on the Learner. Rowley, Mass.

Chomsky, N. (1965), Aspects of the theory of syntax. Cambridge, Mass.

Chomsky, N. (1968), Language and mind. New York.

Christ, H. (1979), "Die Stundentafeln für den neusprachlichen Unterricht in der Entwicklung der letzten beiden Jahrzehnte", Neusprachliche Mitteilungen, 32. Jahrgang, Heft 1, 1-5

Clark, J.L. (1972), Foreign Language Testing: theory and practice. Philadelphia: Center for Curriculum Development.

Clark, J.L.D. (1978), Direct Testing of Speaking Proficiency: Theory and Application. Princeton: Educational Testing Service.

Clark, R. (1977), "Procedures and Computations in the Analysis of Experiments", Davies/Allen, Testing and experimental methods. London, 1977:146-156

Clark, R. (1977), "Statistical Inference", Davies/Allen, Testing and experimental methods. London, 1977:158-174.

Clausing, G. & Lange, D. (1977), "Cloze Testing and Language", Paper presented at the Research Seminar in Foreign and Second Language Learning, University of Southern California, Febr., 1977.

Clauß, G. & Ebner, H. (1970), Grundlagen der Statistik für Psychologen, Päadagogen und Soziologen. Frankfurt a.M. und Zürich, 1970.

Coe, N. (1980),"A closer look at Cloze." Modern English Teacher, Vol.6, Nr. 6, London, 1979, 14-17

Cooper, F. (1972), "How is Language Conveyed by Speech?", F.J. Kavanagh & J.G. Mattingly, Language by Ear and by Eye: The Relationships between Speech and Reading. Cambridge, Mass., 1972. 25-46.

Cripwell, K. (1976), "What is a Cloze Test; how do I use it?" MET, 4/1, 4-7.

Cronbach, L.J. (1970), Essentials of Psychological Testing, New York.

Dale, E. (1948), "Dale list of 3000 familiar words." E. Dale & J.S. Chall, A formula for predicting readability. Educational Research Bulletin, 1948, 27, 37-54.

Darnell, D.K. (1970), "Clozentropy: A Procedure for Testing English Language Proficiency of Foreign Students", Speech Monographs, Vol. XXXVII, March, 1970. 36-46

Davies, A. (^{2}1970), Language Testing Symposium: A Psycholinguistic Approach, London, ^{2}1970.

Davies, A. (1977), "The Construction of Language Tests", Allen/Davies, Testing and experimental methods. London, 1977:38-104.

Doyé, P. (1974), "Verfahren zur Feststellung von Schülerleistungen im Englischunterricht und ihre Eignung für die verschiedenen Lernbereiche", Praxis des neusprachlichen Unterrichts, 21. Jahrgang, 1974. 231-234

Ebbinghaus, H. (1897), "Über eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern", Zeitschrift für Psychologie und Physiologie der Sinnesorgane, 13:401-59.

Englisch Einstufungstest 6+ Form A und B: Entwickelt und herausgegeben von der Projektgruppe "Leistungsmessung in Gesamtschulen" am DIPF. Weinheim, 1972

ETS (1974), Selected Resources in Foreign Language Testing. Princeton: Educational Testing Service

Evans, G. & Haastrup, K. (1976), "Experiments with Cloze Procedure", ELT, Vol. XXXI, Oct. 1976, 35-41

Farhady, H. (1979), "The Disjunctive Fallacy between Discrete-Point and Integrative Tests", TESOL, Vol. 13, No. 3, Sept. 79, 347-357

Farr, R. (1970), Measure and Evaluation of Reading, New York, 1970.

Fehse, K.D. & Praeger, W. (1973), Bibliographie zum Testen. Freiburg.

Fettweiß, B. et al (1966), Hamburger Englischtest HET 6 + ; in der Reihe "Deutsche Schultests", Weinheim, 1966.

Fishman, J.A. & Cooper, R.L. (1978), "The Sociolinguistic Foundations of Language Testing", B. Spolsky, Papers in Applied Linguistics. Advances in Language Testing. Series: 2. Arlington, Virginia, 31-38.

Flesch, R. (1948), "A new readability yardstick", Journal of Applied Psychology. 32 (3), 221-233

Flesch, R. (1951), How to test readability. New York, 1951.

Friedrichs, H. (1975), Englisch H 5, Berlin, ^{2}1975.

Fries, C.C. (1952), The Structure of English. New York.

Fry, E. (1977),"Graph for Estimating Readability - Extended", Journal of Reading, Dec. 1977.

Gaude, P. & Teschner, W. (^{3}1973), Objektivierte Leistungsmessung in der Schule. Einsatz informeller Tests im leistungsdifferenzierten Unterricht. FfM/Berlin/München, ^{3}1973.

Gefen, R. (1979), "An Experiment with Cloze Testing", ELT, 1, 1979, 122-125

George, H.V. (1972), Common Errors in Language Learning, Rowley, Mass.

Gradman, H.L. & Spolsky, B. (1975), "Reduced Redundancy Testing. A Progress Report", J. Randall & B. Spolsky, Testing Language Proficiency, Arlington, Va. 1975, 59-70.

Gregory, P.J.F. (1966), An experimental application of 'cloze' procedure as a diagnostic test of listening comprehension among foreign students. Unpublished Doctoral Dissertation at the University of California, Los Angeles.

Gronlund, N.E. (1968), Constructing Achievement Tests, Englewood Cliffs, New Jersey.

Guilford, J.P. (1956), Fundamental Statistics in Psychology and Education, New York.

Guilford, J.P. (1954), Psychometric Methods, New York.

Halliday, M.A.K. (1970), "Language structure, language function". J. Lyons, New horizons in linguistics, 1970, 140-165

Hanzeli, V.E. (1977), "The Effectiveness of Cloze Tests in Measuring the Competence of Students of French in an Academic Setting," French Review, 50, 6, 865-874, May 77

Harris, D.P. (^{2}1970), "The Linguistics of Language Testing", A. Davies, Language Testing Symposium. A Psycholinguistic Approach. London, ^{2}1970; 36-46.

Harris, D.P. (1969), Testing English as a second language. New York.

Haskell, J.F. (1973), Refining the cloze testing and scoring procedures for use with ESL students. Ed. D. thesis, Columbia University.

Heaton, J.B. (1975), Writing English Language Tests. A practical guide for teachers of English as a second or foreign language. London.

Hecht, K.-H. (1974), Englisch Sekundarstufe I. Unterrichtsplanung und Unterrichtsgestaltung, Bd. I: Grundlagen, Donauwörth

Hecht, K.-H. (1977), Englisch Sekundarstufe I. Bd. II: Unterrichtsgestaltung, Donauwörth.

Hecht, K.-H, & Waas, L. (1980), Englischunterricht konkret. Donauwörth.

Heller, K. (1974), Leistungsbeurteilung in der Schule. Heidelberg.

Heumon, V.A.C. & Bohan, J.E. & Brigham, C.C. (1929), Prognosis Tests in in the Modern Foreign Languages, New York.

Hessisches Institut für Lehrerfortbildung, Protokoll der 12. Arbeitstagung: Leistungsmessung und -beurteilung im kommunikativen Englischunterricht, Fuldatal/Kassel, 1979

Hinofotis, F.A. (1976), An investigation of the concurrent Validity of Cloze Testing as a measure of overall Proficiency in English as a Second Language. Unpublished Doctoral Dissertation, Southern Illinois University.

Hinofotis, F.A. (1979), Cloze Testing: An overview. Revised version of a paper presented at the 1979 CATESOL Conference in Los Angeles. Unpublished.

Hisama , K. (1976), Design and empirical validation of the cloze procedure for measuring language proficiency of non-native speakers. Diss., South Illinois University, Carbondale.

Hisama , K. (1977), A New Direction in Measuring Proficiency in English as a Second Language, New York.

Hörmann, H. (1967), Psychologie der Sprache. Berlin, 1967.

Hosley, D. & Meredith, K. (1979), "Inter- und Intratest Correlates of the TOEFL", TESOL QUARTERLY, Vol. 13. No. 2, June, 1979, 209-218

Hymes, D. (^{3}1976), "On Communicative Competence" in J.B. Pride & J. Holmes, Sociolinguistics, Harmondsworth, ^{3}1976, 269-293

Hymes, D. (1974), Foundations in Sociolinguistics. Philadelphia.

Ingenkamp, K.-H.(1967), "Feststellung der Fremdsprachenleistung und -eignung, in: PRAXIS 3, 1967.
-eignung

Ingenkamp, K.-H. (^{4}1973), Die Fragwürdigkeit der Zensuren-gebung. Weinheim.

Ingenkamp, K.-H. (^{5}1976), Tests in der Schulpraxis. Weinheim.

Ingram, E. (1977), "Basic Concepts in Testing", D.P.B. Allen & A. Davies, Testing and Experimental Methods. 11-37.

Ingram, E. (1978), "The Psycholinguistic Basis", Spolsky, Advances in language testing. Arlimgton: Center for Applied Linguistics.1-15.

Ingram, E. (1977), "Basic Concepts in Testing", Allen/Davies, Testing and experimental methods. London, 1977:11-35

Ingram, E. (^{2}1970), "Attainment and Diagnostic Testing", A. Davies, Language Testing Symposium. A Psycholinguistic Approach. London. 70-98.

Irvine, P. / Atai, P. / Oller, J. (1974), "Cloze, Dictation & the Test of English as a Foreign Language", Language Learning 24, No. 2, Dec. 1974, 245-252.

Jackson, S.H. (o.J.), "Reading Comprehension Questions in Tests of English as a Foreign Language", ohne Angabe des Erscheinungsortes

Jacobovits, L.A. (1970), Foreign Language Learning, Rowley, Mass.

James, C.J. (1978), The Psychology of Cloze in Language Testing, unveröffentlichtes Manuskript eines Vortrages gehalten vor der GAL.

Jones, L.R. & Spolsky, B. (1975), Testing Language Proficiency, Arlington, Va.: Center for Applied Linguistics.

Jongsma, E.R. (1971), The Cloze Procedure: A Survey of the Research. Indiana University, Bloomington, 1971.

Jonz, I. (1976),"Improving on the basic egg: the multiple choice cloze test", Language Learning 26, 2, 1976, 255-265.

Journal of Reading, Dec. 1977

Kahl, P.W. (1971), "Möglichkeiten der Leistungsmessung im Fremdsprachenunterricht", Neusprachliche Mitteilungen, 1971:199.

Kamratowski, J. (1971), "Leistungstests", ENGLISCH 4, 1971:

Kamratowski, J. u.a. (^{2}1973), Informelle Fremdsprachentests für die Schulpraxis, Berlin.

Klare, G. (1974-1975), "Assessing readability", Reading Research Quarterly, Number 1, 62, 63-102.

Klein-Braley, Ch. (1979), "Test Analysis with SPSS", IVS Workpaper, Number 2, November 1979.

Knippschild, H. (^{5}1975), English for you 1, Dortmund.

Krashen, St. (1977), "The monitor model for adult second language performance". M.H. Burt & Dulay & M. Finocchiaro, Viewpoints on English as a Second Language. New York.

Krashen, St. (1977), "Some issues relating to the monitor model". Brown, H.D. & C. Yorio, On TESOL 77. Washington.

Krzyzanowski, H. (1976), "Cloze Tests as Indicators of General Language Proficiency", Studia Anglica Posnaniensia, 1976, 7, 29-43.

Lado, R. (1964), Language Testing, New York. Deutsche Übersetzung: R. Freudenstein, Testen im Sprachunterricht. München, 1971.

Lapkin, S. & Swain, M. (1977), "The Use of English and French Cloze Tests in a Bilingual Education Program Evaluation: Validity and Error Analysis." Language Learning 27 (2). 1977, 279-314.

Lienert, G.-A. (^{2}1967), Testaufbau und Testanalyse. Weinheim.

Lindeman, R.H. (1967), Educational Measurement. Glenview, Ill.

Macht, K. (1975), Lernzielkontrolle im Englischunterricht, Donauwörth.

Madsen, H.S. (1976), "New Alternatives in EFL Exams or 'How to Avoid Selling English Short' ". ELT Journal, XXX, Nr. 2, Jan.76.135-143.

Miller , G. (1956), "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information", Psychological Review, Vol.63, 81-87.

Moller, A. (1975), Validity in proficiency testing, ELT Documents 75/3. London: English Teaching Information Centre. 1975, 5-17.

Munby, J. (1978), Communicative Syllabus Design. London/Cambridge.

Neuner, G. (1977), "Der Reading Comprehension Test, eine neue Form der Abschlußprüfung für die Realschulen", Englisch 1977/2, 42-43.

Nunnally, J.C. (1972), Psychometric Theory, New York, ^{2}1972.

Oakeshott-Taylor, J. (1976), "The Objectivity of Objective Tests: An Alternative View", Die neueren Sprachen 5/1976, 462-471.

Oakeshott-Taylor, J. (1979), "Cloze Procedure and Foreign Language Listening Skills", IRAL XVII/2, 1979, 150-158

O'Brien, M.C. (1972), Testing in Second Language Teaching: New Dimensions. Dublin, University Press.

Oller, J.W. (1971), "Expectancy for Successive Elements", invited paper presented at the TESOL Convention, New Orleans, March 1971.

Oller, J.W.(1978), "Pragmatics and Language Testing", B. Spolsky, Advances in Language Testing. Series 2: Papers in Applied Linguistics. Arlington, Va. 39-59.

Oller, J.W. (1972), "Scoring Methods and Difficulty Levels for Cloze Tests of Proficiency in English as a Second Language", The Modern Language Journal, Vol. LVI, Nr. 3. 151-158

Oller, W.J. (1973), "Cloze Tests of Second Language Proficiency and what they Measure", Language Learning, Vol. 23. 105-118.

Oller, J.W. (1975), Research with Cloze Procedure in Measuring the Proficiency of Non-Native Speakers of English: An Annotated Bibliography. CAL-ERIC/CLL Series on Languages and Linguistics, No. 13. Arlington, Va.

Oller, W.J. (1976), "Evidence for a general language proficiency factor: an expectancy grammar". Die neuren Sprachen. 2. 165-174.

Oller, J.W. (1979), Language Tests at School: a Pragmatic Approach. London.

Oller, J.W. & Bowen, D. & Ton That Dieu & Mason, V. (1972), "Cloze Tests in English, Thai, and Vietnamese: Native and Non-native Performance", Language Learning, Vol. 22, Nov. 1972. 1-15

Oller, J.W. & Conrad, Ch. (1971), "The Cloze Technique and ESL Proficiency", Language Learning, Vol 21, Febr. 1971. 183-196

Oller, J.W. (1971), "A Cloze Test of English Prepositions", TESOL Quarterly 5, 315-326.

Oller, J.W. & Streitf, V. (1975), "A Test of Grammar-Based Expectancies", ELT Journal, Vol. XXX, Nr. 1, Oct. 1975. 25-26.

Oller, J.W. & Perkins, K. (1978), Language in Education: testing the tests. Rowley, Mass.

Oller, J.W. & Richards, (1973), Focus on the learner: pragmatic perspectives for the language teacher. Rowley, Mass.

Osgood, Ch. E. (1953), "The Nature and Measurement of Meaning", Psychological Bulletin 49: 197-237, May, 1952.

Osgood, Ch. E. (1953), Method and Theory in Experimental Psychology, New York/Oxford.

Osieka, H. & Heuser, I. & Messelken, H. (1972), Probleme , Prioritäten, Perspektiven. Leistungskontrolle im Fremdsprachenunterricht. Frankfurt.

Ozete, O. (1974), "Assessing Reading Comprehension in Spanish as a Foreign Language: An Information Process Approach"(Phil. D. Diss., Indiana University)

Palmer, L. & Spolsky, B. (1975), Papers on Language Testing 1967-1974 , TESOL, Washington D.C.

Pauels, W. (1975), "Zur Objektivität 'objektiver Tests'", Die neuren Sprachen, 2/1975. 112-120.

Pelz, M. (1977), Pragmatik und Lernzielbestimmung im Fremdsprachenunterricht. Heidelberg.

Perren, G. (1967), "Testing ability in English as a Second Language", ELT, 1967, Vol. 22:1. 22-29.

Pickett, G.D. (1968), "A Comparison of Translation and Blank Filling as Testing Techniques". ELT, 1968, Vol. 23. 21-26

Piepho, H.-E,(1978), Kommunikative Didaktik des Englischunterrichts, Limburg.

Pike, L.W. (1973), "An evaluation of present and alternative item formats for use in the test of English as a foreign language. (Unpublished MS). Priceton, New Jersey: Educational Testing Service.

Pilliner, A.E.G. (^{2}1970), "Subjective and Objective Testing", A. Davies, Language Testing Symposium. A Psycholinguistic Approach. London, 1970. 19-36.

Pimsleur, P. (1966), Mannual: Language Aptitude Battery. New York.

Pimsleur, P. & Mosberg, L. & Morrison, (1962), "Student Factors in Foreign Language Learning", Modern Language Journal, 46:160-170.

Pimsleur, P. (^{2}1970), "Language Aptitude Testing", A. Davies, Language Testing Symposium. A Psycholinguistic Approach. London. 98-107.

Pimsleur., P. (1966), "Testing Foreign Language Learning", A. Valdman, Trends in Language Teaching. New York, 1966. 175-214.

Porter, D. (1976), "Modified cloze procedure: a more valid reading comprehension test", ELT Journal, XXX, 1976:151-155.

Porter, D. (1978), "Cloze Procedure and Equivalence", Language Learning, Vol. 28, Nr. 2, 1978:333-341.

Pride, J.B. (^{2}1974), The Social Meaning of Language. London, ^{2}1974.

Pumfrey, P.D. (1977), Measuring Reading Abilities. London.

Pynsent, R.B. (1972), Objektive Tests im Englischunterricht der Schule und Universität. Frankfurt/M., 1972.

Quirk, S. (1966), Investigating Linguistic Acceptability. The Hague, 1966.

Radice, F.W. (1976), "The Cloze Procedure as a Teaching Technique", MET 4/1, 1976:6-7.

Radice, F.W. (1978), "Using Cloze Procedure as a Teaching Technique", ELT, Vol. 32, April, 1978:201-205.

Rand, E. (1978), "The Effects of Test Length and Scoring Method on the Precision of Cloze Test Scores", Workpapers in Teaching English as a Second Language, Vol. XII, June 1978. University of California, Los Angeles. 1978, 62-71.

Rankin, E.F. (1970), "The Cloze Procedure - it's Validity and Utility", Fair Measurement: an Evaluation of Reading, New York. 237-253

Red, P.R. (1978), "A Framework for the Assessment of Communicative Competence", Bundesarbeitsgemeinschft Englisch an Gesamtschulen, Leistungsmessung und Leistungsbeurteilung im Kommunikativen Englischunterricht der Sekundarstufe I. Protokoll der 12. Arbeitstagung. Fuldatal: Hessisches Institut für Lehrerfortbildung, 1978:9-26.

Redden, J.E. (1971), Proceedings of the First International Conference on Frontiers in Language Proficiency and Dominance Testing. Carbondale: Southern Illinois University.

Reilly, R.R. (1971)"A note on Clozentropy: a procedure for testing English Language Proficiency of foreign students", Speech Monographs, 1971, 38:350-353.

Riley, P.M. (1973), The Cloze Procedure - a selected Annotated Bibliography. P.N.G.: University of Technology. Lae.

Robinson, R.D. (1972), An Introduction to the Cloze Procedure. An Annotated Bibliography. IRA, Delaware, 1972.

Roeder, P.M. (1974), "Modelle der Differenzierung in Abhängigkeit von Leistungsdimensionen einzelner Schulfächer", P.M. Roeder & K. Treumann, Dimensionen der Schulleistung. Gutachten und Studien der Bildungskommission. Stuttgart, 1974. Bd. 21, 11-131.

Röhr, H. et at (^{2}1970), The Good Companion A, Teil I. Frankfurt/Berlin/München, ^{2}1970.

Salomon, E. (1954),"A Generation of Prognosis Testing", Modern Language Journal. 38:299-303.

Sang & Vollmer, (1978), Allgemeine Sprachfähigkeit und Fremdsprachenerwerb. Zur Struktur von Leistungsdimensionen und linguistischer Kompetenz des Fremdsprachenlernens. Max-Planck- Institut für Bildungsforschung, Berlin, 1978.

Scheibner-Herzig, G. & Pieper, A. (1977), "Beispiel einer Eignungsvorhersage im Fremdsprachenunterricht", H. Christ & H.-E. Piepho, Kongreßdokumentation der 7. Arbeitstagung der Fremdsprachendidaktiker, Gießen 1976. Limburg, 1977:284-287.

Schrand, H. (1976),"Zeitschriftenschau. Schwerpunkt: Testen", Zielsprache Englisch, 4/76:32-33.

Schrand, H. (1970), "Lehrbuchneutraler Leistungstest für das Fach Englisch", Fremdsprachlicher Unterricht, 4, Heft 14, 1970: 2. Beilage.

Schrand, H. (1973), Englisch Einstufungstest 6+ . Weinheim, 1973.

Schrand, H. (1973), Testen . Probleme der objektiven Leistungsmessung im neusprachlichen Unterricht. Berlin, 1973.

Schrand, H. (1975), "Informelle Englischtests", Pädagogische Welt 4, 1975.

Schrand, H. (1976), "Cloze Tests - Anmerkungen und Beispiele", Zielsprache Englisch, Heft 4/76.

Schrand, H. (1976), "Zeitschriftenschau. Schwerpunkt: Testen", Zielsprache Englisch, 4:76:32-33.

Schröder, H. (1976), Leistungsmessung und Schülerbeurteilung. Stuttgart, 1974, ^{2}1976.

Schröter, G. (1977), Zensurengebung. Allgemeine und fachspezifische Probleme. Kastellaun/Hunsrück, 1977.

Shannon, C.E. & Weaver, W. (1963), The Mathematical Theory of Communication. Urbana, Illinois: University of Illinois Press, 1949 (1963).

Shohamy, E. (1978), Investigation of the Concurrent Validity of the Oral Interview with Cloze Procedure for Measuring Proficiency in Hebrew as a Second Language. University of Minnesota. Doctoral Dissertation, June, 1978.

Spolsky, B. (1971), "Reduced Redundancy as a Language Testing Tool", G.E. Perren & J.L. Medd, Applications of Linguistics. Selected Papers of the Second International Congress of Applied Linguistics, Cambridge 1969. Cambridge, 1971.

Spolsky, B. (1978),"Introduction: Linguists and Language Testers", B. Spolsky, Advances in Language Testing. Series: 2. Approaches to Language Testing. Arlington. Va., V-X

Spolsky, B. (1968), "What does it mean to know a language: How do you get someone to perform his competence?" Richards, . & J-W. Oller, Focus on the learner: pragmatic perspectives for the language teacher. Rowley, Mass.(1973) 164-176.

Spolsky, B. (1968), "Language Testing - The Problem of Validation", TESOL 1968:88-94.

Spolsky, B. Current Trends in Language Testing. to appear.

Spolsky, B. (1978), Advances in Language Testing - Approaches to Language Testing. Arlington, Va., 1978.

Spolsky, B. (1979), Advances in Language Testing. Some Major Tests. Arlington, Va., 1979.

Steltmann, K. (1975), Englisch-Tests zur Leistungsmessung und Diagnose, Dissertation. Bonn, 1975.

Steltmann, K. (1978), Faktoren der Fremdsprachenleistung. Kastellaun/Hunsrück, 1978.

Stevenson, D.K. (1978), Face Validity and Loss of Faith: Some Effects of Recent Cloze Research on Traditional Views of Language Proficiency. Paper presented at the GAL Tagung, Mainz 1978.

Stevenson, D.K. (1979), Beyond Faith and Face Validity: the multitrait-multimethod matrix and the convergent and discriminant validity of oral proficiency tests. Shortened version of the paper presented at the 1979 Annual TESOL Convention in Boston.

Stevenson, D.K. (1979), A Selected List of Basic Resources in Foreign Language Testing. IUS Workpaper, Number 1 - June 1979. Essen/Duisburg.

Stubbs, J.B. & Tucker, R. (1974), "The Cloze Test as a Measure of English Proficiency", The Modern Language Journal, Vol. LVIII, 1974:239-241.

Taylor, W. (1953), "Cloze Procedure: a New Tool for Measuring Readability", Journalism Quarterly 30, 415-433.

Templeton, H.A. (1977), A new technique for measuring listening comprehension", ELT, 1977, 31 (4):292-299.

Upshur, J.A. (1973), "Productive Communication Testing: A progress report", J.W. Oller & J.C. Richards, Focus on the Learner: pragmatic perspectives for the language teacher. Rowley, Mass. 1973. 200-213.

Upshur, J.A. & Fata, J. (1968), Problems in Foreign Language Testing: Language Learning Special Issue,Number 3.

Valette, R.M. (^{2}1977), Modern Language Testing. New York, 1967, ^{2}1977. Deutsche Übersetzung: Tests im Fremdsprachenunterricht von Kahl, P.W., Berlin, 1972.

Valette, R.M. (1972), Modern Language Performance Objectives and Individualization. Harcourt, 1972.

Waas, L. (1975), "Leistungskontrolle im Englischunterricht", Pädagogische Welt, 4, 1975.

Wainman, H. (1979), "Cloze Testing of Second Language Learners", ELT, 1979: 126-132.

Weaver, W.W. & Kingston, A.J. (1963), "A factor analysis of the cloze procedure and other measures of reading and language ability". Journal of Communication, Vol. 13, 252-262.

Wechsler, D. (1958), The Measurement and Appraisal of Adu lt Intelligence. Baltimore, 1958.

Wijnstra, J-M. & van Wageningen, N. (1974), The cloze procedure as a measure of first and second language proficiency. Unpublished MS.

Woods, W.A. (1970, 1972), "Transition Network Grammars for Natural Language Analysis" and "An Experimental Parsing System for Transition Network Grammars".

Workpapers in Teaching English as a Second Language, Volume XII, June, 1978. University of California, Los Angeles.

VERZEICHNIS DER TAFELN UND ABBILDUNGEN

A) VERZEICHNIS DER TAFELN:

TAFEL 1 — Anwendung und Forschungsbereich des Cloze-Verfahrens nach James

TAFEL 2 — kontrastive Darstellung der 'differentia specifica' zur Darstellung der Terminologie des Cloze Tests

TAFEL 3 — Darstellung der verschiedenen Validitätsarten in Beziehung zu einem Kriterium nach Davies,[2] 1970:9

TAFEL 4 — Chart of linguistic competence to be filled out for a language system, aus Davies, 1968:54,55

TAFEL 5 — Verteilung der Stundenzahl des Englischunterrichts an den diversen Schularten pro Woche

TAFEL 6 — Mean scores of subjects on cloze test, aus Oller/Conrad, 1971:189

TAFEL 7 — Pearson product moment correlation matrix for TOEFL part score, the cloze test scored by exact word and acceptable word methods, and two dictations, aus Irvine/Atai/Oller, 1974:249

TAFEL 8 — Zur Validierung vorgenommene Korrelationsberechnungen zwischen den CT- und dem Clozentropy-Verfahren und den TOEFL Subtests

TAFEL 9 — Reliabilitätsfaktoren und Testwerte im Experiment von Hisama

TAFEL 10 — Korrelationen mit dem Ganzen und den Subtests aus Krzyzanowski, 1976:34)

TAFEL 11 Correlation between test results and the teachers' classification, aus Krzyzanowsky, 1977:36

TAFEL 12 Übersicht über die ermittelten Werte des Experiments von Hinofotis, 1976:94

TAFEL 13 Werte der Validierung aus Hinofotis, 1976:106

TAFEL 14 Gefundene Korrelationen mit den Subtests, aus Hinofotis, 1976:104

TAFEL 15 Pearson Product Moment correlations of cloze scores with ELBA and dictation by test, aus Alderson; 1978:278

TAFEL 16 a) Studien zum Standardmodell des Cloze Tests als Meßinstrument der Fremdsprachenkompetenz: Reliabilitätskoeffizienten

TAFEL 16 b) Studien zum Standardmodell des Cloze Tests als Meßinstrument der Fremdsprachenkompetenz : Validitätskriterien

TAFEL 17 Correlation of ELBA Total with various cloze tests, aus Alderson, 1978:276

TAFEL 18 Modell der Language Comprehension von Fillmore

TAFEL 19 Aufstellung der verschiedenen Probandengruppen, gegliedert nach Schulart/Schülerjahrgang, Schule/Klasse, Probandenzahl

TAFEL 20 Durchführungsdaten von drei Cloze Tests in den Jahrgangsstufen 5-9 verschiedener Haupt- und Realschulen sowie Gymnasien

TAFEL 21 Arithmetische Mittel der Einzeltests und der Gesamttestbatterie in allen Klassen und Schularten

TAFEL 22 Leistungsprofile

TAFEL 23 Mittlere Schwierigkeitsgrade von drei CTs und der Kombination dieser drei Tests bei den Jahrgangsstufen 5-9 aller drei Schularten

TAFEL 24 Verhalten der Jahrgangsstufen 5-9 verschiedener Schularten bei drei CTs verschiedener Textschwierigkeit

TAFEL 25 Streuwerte (Range und Standardabweichung) bei drei CTs und der Kombination der drei Tests für die Schülerjahrgänge 5-9 aller Schularten

TAFEL 26 Darstellung der Standardabweichungen und Streubreiten ('Ranges') des CTs "Who is Who?" bei den Klassenstufen 5-9 aller Schularten

TAFEL 27 Darstellung der Standardabweichungen und Streubreiten des CTs "Pat at the Grocer's" bei den Klassenstufen 5-9 von HS, RS und Gym.

TAFEL 28 Darstellung der Standardabweichungen und Streubreiten des CTs "Tim's First Drink" bei den Jahrgängen 5-9 von HS, RS und Gym.

TAFEL 29 Darstellung der Standardabweichungen und Streubreiten der CTs I, II und III bei den Klassenstufen 5-9 aller Schularten

TAFEL 30 Arithmetisches Mittel der Jahresfortgangsnote im Fach Englisch und der Einschätzung zur Beurteilung der Fremdsprachenkompetenz von Schülern der Klassen 5-9 aller Schularten

TAFEL 31 Reliabilitätskoeffizienten nach Spearman-Brown und Kristof für drei verschiedene CTs und deren Kombination bei den Klassen 5-9 aller Schularten

TAFEL 32 Reliabilitätskoeffizient berechnet nach Spearman-Brown für 100 Aufgaben für drei CTs mit 25 Aufgaben und deren Kombination in den Jahrgangsstufen 5-9 aller Schularten

TAFEL 33 Korrelationskoeffizienten zwischen den Rohpunktwerten von drei Tests und deren Kombination und zwei Validierungskriterien für die Jahrgangsstufen 5-9 bei allen Schularten

TAFEL 34 Interkorrelation der drei CTs bei den Jahrgangsstufen 5-9 von HS, RS und Gym.

TAFEL 35 Zur Interdependenz der statistischen Werte bei CT I und der Kombination dieser Tests

TAFEL 36 Zur Interdependenz der statistischen Werte bei CT II und der Kombination dieser Tests

TAFEL 37 Zur Interdependenz der statistischen Werte bei CT III und der Kombination dieser drei Tests der Jahrgangsstufen 5-9 aller Schularten

TAFEL 38 Zur Interdependenz der statistischen Werte bei der Kombination der drei Tests CT I, II und III der Jahrgangsstufen 5-9 aller Schularten

TAFEL 39 Gesamtübersicht der statistsichen Ergebnisse der Durchführung einer Testbatterie bei Schülern der Jahrgangsstufen 5-9 der Haupt- und Realschulen und Gymnasien

B) VERZEICHNIS DER ABBILDUNGEN:

ABB I — Die Reliabilität einer Testform A in Korrelation zu Testform B dargestellt in einem Korrelationsdiagramm, nach Lienert, ^{2}1967:238)

ABB II — Darstellung verschiedener Korrelationen (Beispiel von r = 0, 1,-1 und 0,9) in einer selbsterstellten Graphik

ABB III — Gesamtmodell der "Sprachlichen Kompetenz", in Abänderung entnommen aus Hecht, 1980:14

ABB IV — Profiles of di fferent language groups, aus Hisama, 1976: 116

ABB V — Reliabilitätskurve in Abhängigkeit der Itemanzahl, aus Rand; 1978:69

C) VERZEICHNIS DER MATRICES:

MATRIX I — Übersicht über verschiedene Testarten und deren Autoren, Harris, Valette, Heaton, Davies, Clark und Lado

MATRIX II — Teilbereiche der Intentionalität von Fremdsprachentests

Lebenslauf

Persönliche Daten:

Name:	Klaus Dieter Wolfgang Balik
Adresse:	8032 Lochham/München, Aubingerstraße 38
Familienstand:	verheiratet mit Patricia Young-Balik
Geburtstag:	17. Mai 1944
Geburtsort:	Garmisch-Partenkirchen
Kinder:	Ein Sohn, Nicolas, geboren am 17.10.1979

Schulbildung:

1950 - 1954	Volksschule Gräfelfing
1954 - 1963	Realgymnasium Gräfelfing (Abitur)
1963 - 1967	Universität München / Pädagogische Hochschule Studium der Erziehungswissenschaften I. Staatsprüfung für das Lehramt an Volksschulen
1965	University of Cambridge, Local Examination Syndicate München: Certificate of Proficiency in English
1970	II. Staatsprüfung für das Lehramt an Volksschulen
1970 - 1971	Ball State University, Muncie Indiana, Study of Teaching English as a Foreign Language
1971 - 1973	University of Southern California in Los Angeles: Study of Teaching English as a Foreign Language
1972	Certificate of Teaching English as aForeign Language USC
1973	M. S. Ed. in Teaching English as a Foreign Language USC
1975 - 1980	Universität München Promotionsstudium
Feb. 1981	Promotion an der Universität München mit der Dissertation "Der Cloze Test als Meßinstrument der Fremdsprachenkom - petenz von Schülern mit Deutsch als Muttersprache (eine empirische Studie) "